INTEGRATING NON-TOPICAL ASPECTS INTO INFORMATION RETRIEVAL

A Dissertation Presented

by

ELIF AKTOLGA

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

INTEGRATING NON-TOPICAL ASPECTS INTO INFORMATION RETRIEVAL

A Dissertation Presented

by

ELIF AKTOLGA

Approved as to style and content by:

James Allan, Chair

W. Bruce Croft, Member

David A. Smith, Member

Lisa Ballesteros, Member

Lori A. Clarke, Chair
School of Computer Science

To all the girls who strive for success.

ACKNOWLEDGMENTS

"No problem can be solved from the same level of consciousness that created it – Albert Einstein."

The PhD experience has been a time of personal and professional growth and learning for me. The experiences I went through on this journey have shaped me in many different ways. I am here today where I am because of many reasons:

- I thank each one of the obstacles and problems I have faced in life so far for challenging and shaping me in manifold ways;
- I thank each one of the people I have encountered on my journey, who have helped me flesh out an idea, encouraged me to not give up, engaged with me in interesting conversations, stood by me and offered uplifting advice, helped me out when I needed something, or simply cooked great food for me. At this point, I would like to name them:
  - my grandmother: I dedicate this thesis to her and to all the girls and women in the world. Education is the key to success and independence;
  - my aunt who has fought very hard with cancer for the past four years – you will be greatly missed;
  - my mother, father and my sister whom I can always count on;
  - my remaining extended family across three continents;
  - my advisor James Allan for teaching me all about IR, guiding me on my way to become a better researcher, and for encouraging me to both work

independently and to collaborate with others. Thanks for giving me the flexibility to work the way I did;

– my thesis committee members Bruce Croft, David Smith, and Lisa Ballesteros for their time and helpful comments on this work;

– my internship mentors Alpa Jain and Emre Velipasaoglu during my time at Yahoo, and Irene Ros and Yannick Assogba at IBM;

– my undergraduate and graduate mentors and advisors Ute Schmid and Des Watson;

– Thank you, David Fisher, for your help, not only with Indri and Galago but also beyond;

– Thanks Andrew McCallum and Deepak Ganesan for your help and support early during my PhD;

– the CIIR staff: Kate Moruzzi, Dan Parker, Jean Joyce, and Glenn Stowell;

– special thanks to Leeanne Leclerc for always having an answer to our questions;

– the IPO: in particular Patricia Vokbus for always having a quick solution;

– to all the CIIR lab students, present and past: Niranjan, Michael and Marina, Jie, Ethem, Marc, Jeff, Van, Laura, Shiri, Henry, Ashish, Mostafa, Weize, Xiaobing, Zeki, Xing, Jinyoung, Kriste, CJ, Tamsin, Pranav, Jae-Hyun, Tiger, Josh, Giridhar, Matt, and Jangwon;

– the many rescued rabbits I temp-fostered: these taught me patience and that there is always an upward climb no matter how bad the situation is;

– my friends Jean, Julia, Sebastian, Christine, Katrin, and Ole;

– Thanks for the support and for being there, Devesh and his family;

– Thanks to the 'Lady in Grey' in my life, Tontosh, for all the funny, furry, and enjoyable moments;

– and finally: Audible for making my commutes to Amherst more enjoyable, PhD comics for the humor, and Fitness Blender for the much needed exercise!

Thanks, dear reader, for listening. I hope you will find the ideas presented in this thesis useful and enlightening.

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 by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023, in part by NSF grant #IIS-0910884, and in part by UpToDate. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

ABSTRACT

INTEGRATING NON-TOPICAL ASPECTS INTO INFORMATION RETRIEVAL

MAY 2014

ELIF AKTOLGA

BSc, UNIVERSITY OF OSNABRÜCK

MRes, UNIVERSITY OF SUSSEX

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor James Allan

When users investigate a topic, they are often interested in results that are not just relevant, but also strongly opinionated or covering a range of times. To get such results, users are forced to formulate ambiguous, complex, or longer queries. Commonly this becomes a burden, since users need to issue several queries with reformulations if initial search results are not completely satisfactory. In this thesis, we focus on those two non-topical dimensions: opinionatedness and time. We develop measures for quantifying them in documents and incorporate them into search results.

For improving search results with respect to non-topical dimensions, we use diversification approaches. To achieve controlled variety in results, our methods are integrated with a general bias framework, which seamlessly unifies extreme biases for each dimension. Results can be diversified across a single or multiple non-topical dimensions. Our experiments are performed on the TREC Blog Track.

As a result of this research, we can determine how temporal or opinionated a unit of text is. By means of diversification we provide a retrieval framework to users with which they can more easily find different kinds of opinionated or temporal results with only one submitted query. The burden of analyzing pre-existing biases for a query and discovering times at which important events happened is fully carried by the system.

As opposed to prior work in this area, pre-existing biases in search results are analyzed, and diversification is performed in a controlled manner for each dimension. We show how to combine several dimensions with individual biases for each, while also presenting approaches to time and sentiment diversification. The insights from this work will be very valuable for next generation search engines and retrieval systems.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

PROBLEM STATEMENT

1.1 Introduction

Many web search engines work well for precision-oriented tasks: for most queries they can successfully find the right documents on the first results page. However, particularly when users are investigating a topic where the answer is not contained in a single document, more results about the topic are required. Instead of burdening the user with an overwhelming amount of results to look at and prune, it would be useful to filter or post-process the retrieved results to favor certain characteristics, such as time and opinionatedness. These characteristics are what we call *dimensions with non-topical aspects* since they are not part of the topical content, but rather pose a high-level restriction on the results (Huang & Croft, 2009; Demartini & Siersdorfer, 2010; Demartini, 2011). Expressing such non-topical preferences as part of the query often does not fully satisfy the information need. For example, if we wanted opinionated or provocative search results for the query 'global warming', then reformulations like 'pro global warming', 'indifferent opinion global warming' or 'inclined towards protecting the environment global warming' would mostly yield unsatisfactory results. Further, it is often not clear how to express non-topical preferences as part of the query to achieve good results. The queries can become ambiguous, complex, or long, each of which makes it less likely that the results will be satisfactory. This is a knowledge gap problem since often the user does not know how to query to fulfill her information need (Verberne, 2011) — knowing which related information is relevant and would complete the picture or help retrieve the required information.

Users may have varying search intents, from obtaining simple informative results about topics such as 'algorithms', 'weather in Amherst', or 'earthquake in Asia'. In such cases, users rather expect objective results about the topic. They may also be looking for opinions, or provocative results in order to better inform themselves about controversial topics such as 'homosexuality', 'abortion', or 'cloning'. Kacimi and Gamper (2011) define two types of *controversial queries*: *informative queries* and *debatable queries*. Informative queries are from controversial topics such as 'child abuse' and 'racism' on which *a single sentiment* is predominant. Queries such as 'homosexuality' etc. mentioned above are classified as *debatable queries*. The search results for these queries express many different sentiments, which makes the topic debatable. Whether informative or debatable, users often search with the most general form of the query such as 'global warming' in order to obtain more relevant results. After some initial reading through the results they infer some aspects and keywords from the documents to further search with and reformulate the query. Such an approach however does not necessarily guarantee high recall about the topic, even if repeatedly applied. It just biases the results towards what the user has seen and inferred from the initial results.

Search engines currently try to address this problem by topically diversifying search results for the user. One approach is to use a query's search intents to then diversify the results across those intent categories (Agrawal et al., 2009). This rather improves recall in a general sense. There is no control of the direction in which results are being diversified, nor of whether the results will indeed suitably address a *non-topical* user preference. Therefore, for addressing non-topical preferences of one or several dimensions, a different approach is required: controlling the overall bias in the results while diversifying over query aspects of a non-topical dimension. For this, we tightly integrate diversification frameworks with a general bias framework.

Before thinking about dimensions with non-topical aspects, what does *diversification* actually mean? In prior work, topical diversification refers to minimizing redundancy among search results while maximizing the number of topical arguments discussed (Carbonell & Goldstein, 1998; C. L. Clarke et al., 2008; Agrawal et al., 2009; R. L. Santos et al., 2010a). So while as many results as possible should be obtained about the topic, each result should cover a new or previously unseen aspect of the topic. This typically requires an incremental model that retrieves results and boosts those containing new and different information to what has already been included in the reranked list. In these prior works, it is implicitly assumed that a set of discrete criteria is available across which diversification is performed. These criteria may be query intents, unique novel arguments, or other countable criteria. What does it mean for search results to be diverse? That each aspect is covered at least once, or that each aspect is covered equally many times as others, or maybe even that the aspects are covered according to some distribution? Prior research typically assumes an equal distribution across all aspects for diversification (R. L. Santos et al., 2010a, 2010b, 2011). The only prior work that differs in this point is that of Dang and Croft (2012): the proportionality-based models can diversify search results according to a specified distribution. However, due to lack of suitable experimental data, the models were tested with equal aspect distributions only.

Non-topical diversity is a new area of research. Publications on one dimension with non-topical aspects – opinionatedness – are very recent, however they also employ equal aspect diversification (Demartini & Siersdorfer, 2010; Demartini, 2011; Kacimi & Gamper, 2011). In this thesis, we relax the equal aspect distribution assumption: given a query and its non-topical aspects for a dimension, we estimate the 'true distribution' of these query aspects from a reliable data source. For example, if the query is 'global warming', and the dimension we are interested in is 'sentiments', typical aspects are 'positive', 'negative', and 'neutral'. Analyzing relevant data about

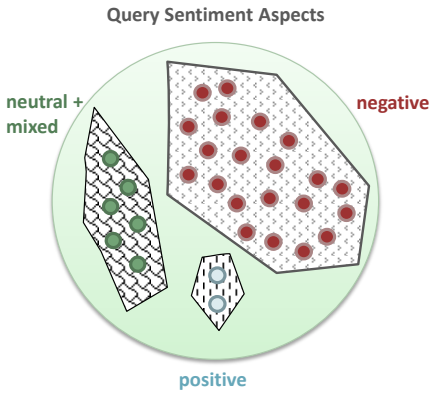

Figure 1.1. Query Sentiment Aspects: Dots represent units of information grouped by sentiments for this query: the obtained query sentiment aspects distribution is used for sentiment diversification.

global warming for sentiments yields the pre-existing bias or query distribution, such as shown in Figure 1.1: for example, 70% negative, 20% neutral and mixed, and 10% positive. We can then use this target query sentiment aspects distribution or *bias* to diversify the search results in different ways: it can be employed 'as is' to emphasize the actual or pre-existing bias for the dimension in search results. So in a ranked list of ten documents for global warming, roughly 7 negative documents, 2 neutral documents, and 1 positive document can be included to mirror this trend. Alternatively, we can change the target query distribution to emphasize minority aspects that are underrepresented in this bias, such as the 'positive' aspect: instead of showing only 1 positive document, we can include more positive documents and reduce the quantities of the other sentiments accordingly. As another alternative, we can ignore the query distribution and perform equal diversification as a target similar to prior work. Finally, we can mix these distributions to achieve a combination that lies on the continuum. Such approaches strongly call for a general bias framework, which is one of the contributions in this thesis.

The benefits of providing such a retrieval framework for diversification are manifold: users only need to specify a query, what kind of bias they desire for each

dimension, and how these dimensions shall be combined, in order to find different kinds of opinionated or temporal results about the query. The burden of analyzing pre-existing biases for a query and discovering times at which important events happened is fully carried by the system.

In this thesis, we explicitly focus on dimensions with non-topical aspects only. We assume that non-topical aspects are crisply defined with clear divisions between different aspects. For application to the topical dimension, our proposed bias framework would have to be carefully considered taking this into mind. Further, note the following facts: if a reranked list is diverse with respect to a non-topical dimension, it may also be topically diverse, but the reverse does not necessarily follow: search results may be novel and unique but not opinionated or temporal. The two forms of diversity – topical and non-topical – can be explicitly or implicitly combined, but the focus in this thesis is to explicitly model the diversification for dimensions with non-topical aspects only.

1.2 Dimensions with Non-Topical Aspects

In this section we informally introduce the two dimensions with non-topical aspects, *opinionatedness* and *time*, which we will deal with further in this thesis. Then, we represent *interestingness* as a label referring to the combination of multiple dimensions with their non-topical aspects. We clarify the meaning and scope of each dimension, and how they apply in a retrieval environment in practice. This is useful background material for the more formal work in Chapters 3, 4, and 5.

1.2.1 Opinionatedness

A unit of text may be written in a subjective or objective tone. The user is expected not to react emotionally to objective text, but she may feel angry, annoyed, or happy, proud, glad etc. after reading subjective material. This is because she

may disagree or agree with the author on the content. There are certain topics about which it is difficult to write objective content because by nature they are what we call *controversial*. There is no agreed upon truth regarding the topic, so such documents are bound to be subjective or opinionated. We believe there are more fine-grained aspects to opinionatedness: documents from a controversial topic are not only subjective, but also provocative if the degree of subjectivity is high. The more subjective a document from such a topic is, the more provocative it is. Therefore, the provocativeness of a topic measures the degree of subjectivity of the topic, which weighs the quantity of subjective versus objective content. For *measuring* the provocativeness of a topic, the actual distribution of sentiments is unimportant – only the ratio of subjective versus objective content on the topic matters. When considering the diversity of sentiments in a topic however, the distribution and variety of provocativeness in a topic becomes important. Thinking about these concepts in a user-independent manner we can say that the more likely a person is to have certain strong feelings about a topic, the more provocative or opinionated it is. On the contrary, if only a little subjective material exists, then the likelihood of a person feeling strongly in a particular direction about the topic (such as being annoyed) is smaller.

By boosting provocative or opinionated documents in search results, we can provide users with a richer and more diverse source of information on controversial topics. This allows them to discern various opinions and sentiments on the topic, such as is shown in Figure 1.2: when searching for 'Mahmoud Ahmadinejad', instead of only reading objective news about the subject, the reader may be interested in opinionated search results. One matter is controlling the direction of the sentiments. To capture this, in Chapter 3 we define the 'balance' measure, which characterizes the direction of sentiments for a topic. Another matter is showing which provocative results stand out from the 'majority sentiment'. Different sentiments can be shown in equal quan-

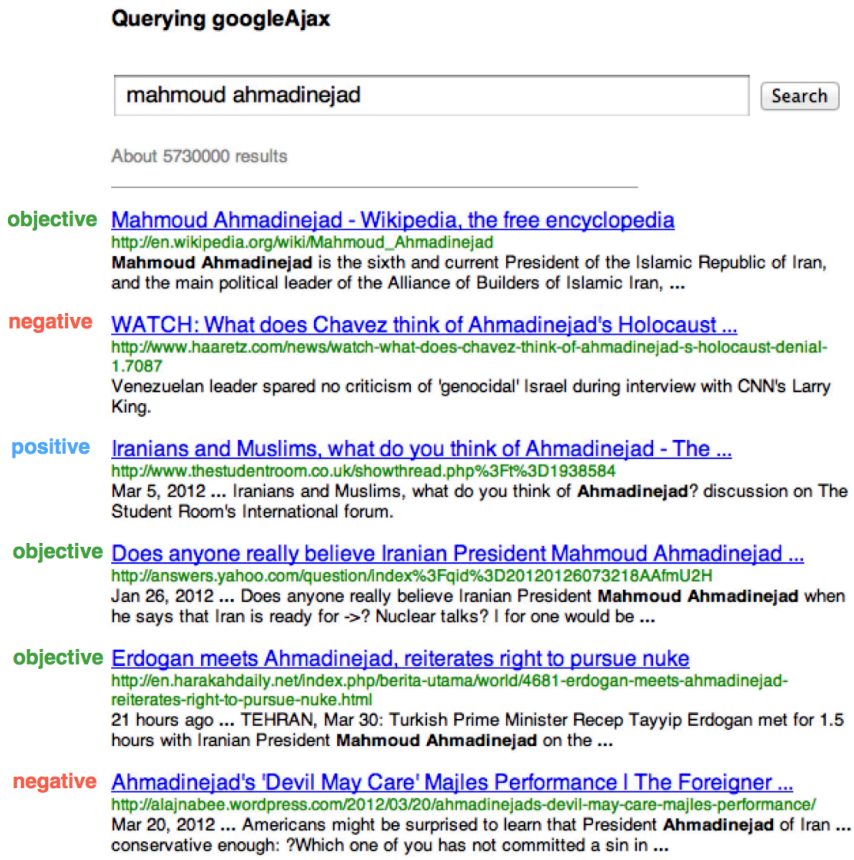

Figure 1.2. An example of diversifying search results for the query 'mahmoud ahmadinejad' to include strongly opinionated documents in highly ranked results.

tity, or the results can be shown in favor of the majority sentiment – or against it. There are many interesting directions here.

1.2.2 Time

We define a temporally relevant document as one exhibiting relevant temporal aspects to the topic or query with which it is retrieved. A document may be associated with many temporal aspects, such as a publication date, revision date, or time references within document content. Web search systems primarily use the publication date of a document to make decisions about its *freshness* or *recency*: for ordinary search intents, users prefer more recently written documents because they want to have the newest information about the topic. However, in situations such as

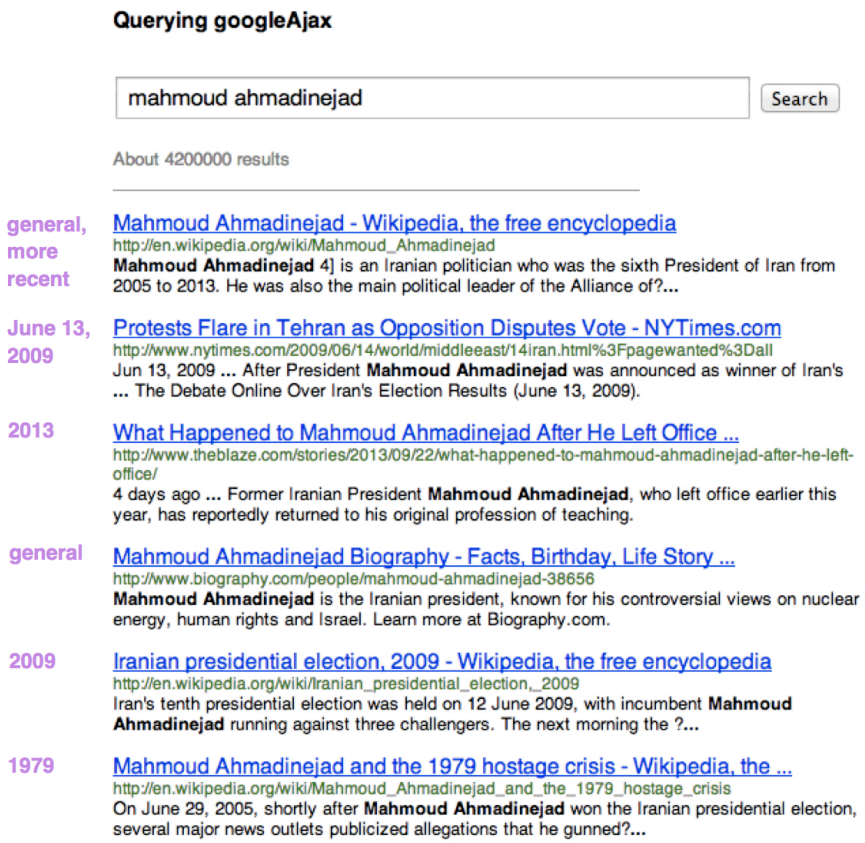

Figure 1.3. An example of diversifying search results for the query 'mahmoud ahmadinejad' to include information about the person from different times.

a literature review, the newest documents may not exclusively be the most valuable ones: older but relevant documents about events related to the topic may also contain useful information. Hence, when working with time, we are particularly interested in (1) detecting important times for the topic: these can be more recent and more distant time ranges. We prefer *time ranges* rather than dates because they can span an arbitrary amount of time, capturing temporal relevance at different granularities; (2) detecting and boosting documents talking about these important times for the topic. These two characteristics are also employed in the definition of a time measure in Section 4.3.

An initially retrieved list of search results for a query without temporal cues may favor more recent search results such as Figure 1.2, which was compiled from web search results from 2012. However, we can explicitly diversify for time and boost documents talking about important events related to the topic such as shown in Figure 1.3. Apart from delivering relevant content for the literature review, the user can discern important times from the documents without spending time to discover those first. In Figure 1.3, these are for instance: the presidential elections in Iran from 2009 and the protests that followed, the 1979 hostage crisis with which Mahmoud Ahmadinejad was claimed to be associated, and other times that can be inferred from a biography. In order to manually retrieve such documents, the user would first have to know about popular or spiking events in the past with the politician, and she would also have to query the search engine several times with time indicators such as "in 2009" or " in the early 2000s" etc. Note that for the presentation of the results it is not required to be fixated on "the most important times" for a topic. Once we have a ranked set of time ranges, the user can choose to favor them in different ways, such as boosting most or least popular times for the topic for greater flexibility during searching. This task requires us to work with time references mentioned *within* documents, which is less typical in the information retrieval literature, where document publication dates are typically preferred.

1.2.3 Interestingness

Given a ranked list of documents and a query from which we started, which documents are "interesting"? If several of them are interesting, which ones are more interesting and why? Interestingness seems to be a user-dependent, subjective criterion, but just like relevance it can be studied in a user-independent way. This is how we interpret interestingness in our work.

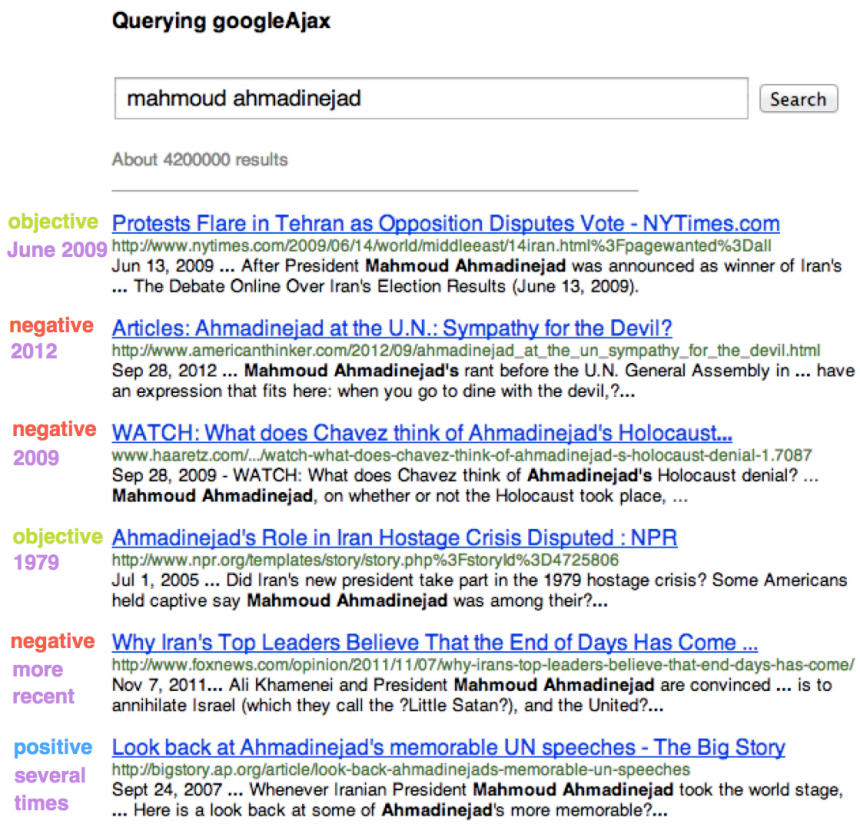

Figure 1.4. An example of diversifying search results for the query 'mahmoud ahmadinejad' to include opinionated and neutral results about the person from different times.

In this thesis, a document is defined as more "interesting" than another if it more successfully addresses certain non-topical aspects of *multiple* dimensions. For this, given a query and a set of dimensions with non-topical aspects, such as opinionatedness, time, geography, or others, we can determine how important each dimension is to a particular query. For example, for a highly controversial query the opinionatedness dimension may be more important than time. Additionally, we also need to consider how well each document fulfills the non-topical query aspects of each dimension. This decision is determined based on the user's desired outcome (for instance, highly opinionated documents from important times for the query). Given a set of preferences by the user and these two characteristics, i.e., (1) the importance of each

dimension to the query, and (2) how well each document addresses each dimension, we can then rank the documents in order of interestingness after retrieval. This is the essence of our interestingness measure defined in Section 5.3.

The term 'interestingness' is a suitable label for summarizing multiple dimensions with non-topical aspects, since it signifies that the interesting object (here: documents) stands out from the pool without explicitly revealing what the distinguishing features are. In this thesis, we know that these distinguishing features are non-topical aspects of several dimensions. We could potentially analyze many such dimensions, but in this work we focus on the interplay between opinionatedness and time only.

Imagine a user who is interested in opinionated results about 'Mahmoud Ahmadinejad', however the current result set rather contains neutral information about very recent events or even rather timeless information. In this case, the user would clearly benefit from a results set with more opinionated content which is also diversified over events in the past where something important happened with Mahmoud Ahmadinejad. An example diversified list is shown in Figure 1.4, which is a blend of Figure 1.2 and 1.3, and also includes a few new documents: the last one is a subjective document about Mahmoud Ahmadinejad's memorable UN speeches, and the one before that titled "Why Iran's Top Leaders Believe That the End of Days Has Come" is another subjective document from 2011. By simultaneously diversifying for multiple dimensions we heavily simplify the task of finding these interesting documents for the user, corresponding to her preferences.

Therefore, in this thesis we describe approaches to diversify for sentiments and time in one step without the user having to do anything manually other than querying for the politician's name. Including interesting documents in search results will allow users to detect aspects of the query they were not aware of earlier, or allow them to view the query's topic from a different perspective. Therefore, such results are

meant to complete the picture and to provide a better understanding of the topic while drawing the user's attention.

1.3 Contributions

The contributions in this thesis are:

1. **Measures.** We define three measures, provocativeness, balance, and average sentiment of a topic to describe its opinionatedness. These measures operate on a 'topic', i.e., a collection of documents. They compare the amount of subjective versus objective data available for the topic, the balance between positive and negative content, and reveal the overall sentiment for the topic given all the data. We also define time, sentiment, and interestingness measures to rank documents in order of relevance of these dimensions or their combination. These measures are employed as part of diversification models.

2. **Sentiment Diversification.** We extend two existing diversification frameworks, xQuAD (R. L. Santos et al., 2010a) and Diversity by Proportionality (Dang & Croft, 2012), to work with the sentiment dimension and the query sentiment aspects positive, negative, and neutral. We also propose variations to these algorithms: xQuAD is modified in Section 3.2.2.1 to not only use the strength of sentiment (or relevance) scores as originally defined, but to also consider the frequency of such sentiment aspects to control diversity. The xQuAD-like adapted version of the model is referred to as SCS, whereas our alternative version is the SCSF model. In the experiments on the TREC Blog Track, SCSF performs comparably to the proportionality models, significantly enhancing the results over the SDM baseline and SCS depending on sentiment classification accuracies. For the proportionality models, we propose a variation to PM-2 (Dang & Croft, 2012) that adapts the quotient calculation in case

there are not enough documents present in the retrieved list for a certain aspect. This model performs comparably well to PM-2, and improves those results when sentiment classification labels are not perfect. Lastly, we show how to integrate three target biases in the diversification frameworks, which makes the models more flexible towards bias-specific diversification than in their original definitions in prior work.

3. **Temporal Diversification.** We show how to do temporal diversification with times extracted from *within documents* as opposed to using document publication dates as in prior work. This yields a larger number of time interval aspects for diversification, variable for each query, and spanning different amounts of time. The above-mentioned diversification frameworks and algorithms for the sentiment dimension are adapted to the time dimension using the same style of biases. Diversification performance is evaluated across time bins extracted from relevant documents, from which it becomes evident that there is potential for large improvements over the SDM baseline with our diversification models: with perfect time labels, we obtain a maximum relative improvement of 40% in Precision-IA@20 over SDM with the Spike bias. For the Slab bias, the gains over SDM are 80%+ for rank-based measures such as α-NDCG@20, ERR-IA@20, and NRBP. With noisy Wikipedia time aspects and weights, we can improve low Precision-IA@20 for the Slab bias up to 40% by collapsing time intervals and their weights as opposed to not altering anything about the setup.

4. **Diversifying across Multiple Dimensions.** We adapt the above-mentioned diversification frameworks for sentiment and temporal diversification to work with multiple dimensions in a single framework, and therefore show how to simultaneously diversify across several dimensions using different biases for each. We show how time aspects and their weights can be collapsed to yield an average

significant improvement of 7.5% when used with the Slab bias as opposed to not modifying those time aspects. We empirically show the effectiveness of the proportionality and DCSF models over the SDM baseline and DCS with noisy labels by evaluating over aspect bins of each of the dimensions. By means of an example, we also demonstrate why and how diversifying across multiple dimensions is more useful than doing the same over a single dimension.

5. **Bias Framework for Non-Topical Diversification.** We propose a general bias framework that seamlessly integrates the three target biases defined for sentiments and time to work with any dimension with a fixed or variable number of query aspects (that however need to be finite), and present two variations to this framework: for the Outlier or Slab biases, one version inverts the original query dimension distribution, whereas the other version reverts it. We conduct experiments on the TREC Blog Track to evaluate the efficiency of the diversification frameworks and biases given different kinds of settings: noisy versus perfect query aspect labels. For all these experiments, we confirm the enhanced performance of the DCSF and proportionality models from earlier experiments. The DCS model (originally xQuAD) only performs comparably well under perfect labels, but proves unstable with noisy labels. We also simulate the lack of data and the effect of substituting biases for one another: on average over 10% of performance is lost according to several measures for the DCSF and PM-2 models if equal diversification is employed instead of the actual intended biases.

1.4 Outline

In Chapter 2 we survey prior work, which touches the areas of topical diversity, non-topical diversity for dimensions such as sentiments or opinions, information retrieval with temporal or time-related aspects, and interestingness.

In Chapter 3 we first introduce concepts related to the dimension opinionatedness, define several measures such as provocativeness, balance, and average topic sentiment that allow us to quantify a piece of text according to this dimension. Then, we consider how this dimension can be employed in diversification in the form of query sentiment aspects. We present several diversification frameworks, and three different target biases to manipulate the Query Aspects Distribution for sentiments, which are used during diversification. Following the experimental results, we also show an example to demonstrate the effect of non-topical diversification with a specific bias for the sentiment dimension.

In Chapter 4 we focus on the time dimension, extracting time information from within documents. We first clarify issues about how to obtain time information from documents. Wikipedia is a suitable source for obtaining time diversification bins, so we show some analysis with this corpus using our set of queries. We present a time measure in Section 4.3, according to which documents can be ranked in order of relevance to this dimension. Afterwards, we show how the diversification models presented in Chapter 3 can be adapted to the time dimension. Finally, some diversification results are presented for the time dimension with the three extreme target biases from Chapter 3.

In Chapter 5 we define 'interestingness' as a label combining several dimensions with non-topical aspects, and choose two, sentiments and time, as dimensions to focus on in this chapter. By means of a query log we reveal evidence about users being interested in subjective and temporal results. Then, we define interestingness as a measure composed of several dimensions, according to which documents can be ranked. Following this, we present diversification frameworks with several models, which are adapted from the approaches in Chapter 3 to work with multiple dimensions. We also include a generalized bias framework that seamlessly integrates the three extreme target biases introduced in Chapter 3. Following experiments with sen-

timent and time judgments, we look at the results from different angles, and present examples to demonstrate why diversifying with several dimensions can prove more useful than diversifying individually for each.

In Chapter 6 we recapitulate the achievements and discoveries of this work, and present directions for future work in various related areas.

CHAPTER 2

RELATED WORK

We have touched aspects of related work in the previous chapter as part of the introduction to diversity and dimensions with non-topical aspects. In this chapter, we detail related work in the areas of novelty, topical diversity, and non-topical diversity more comprehensively. For non-topical diversity, we particularly focus on work in related non-topical areas: opinion retrieval, sentiment analysis, temporal aspects, and interestingness. In the next section, we first clarify the various interpretations of the term 'topic', and then we provide an introduction with the necessary background for understanding the various concepts involved.

2.1 Topic

The term topic is widely used in information retrieval to mean a number of things. Most commonly it refers to the subject of discourse, so documents discuss topics, people are interested in topics, and queries relate to topics. Within the TREC evaluations [1], a 'topic' is a written description of what someone is interested in finding information about. Such a 'TREC topic' is usually an instance consisting of a topic title, a topic description, and a narrative. It often includes some sample queries that someone interested in that topic might consider. Within language modeling, a multinomial distribution over words is generated which can be used to describe the 'topic' or contents of a document. Similarly, in topic modeling, which is a technique for

[1] http://trec.nist.gov/

discovering 'topics' that occur in a collection of documents, a topic can be any relevant abstract concept mentioned in the documents. As discussed later in Section 2.3, Topic Detection and Tracking (TDT) is another area that studies topics, where the notion of a topic extends to 'events' or 'activities'.

Within this thesis, we use the term 'topic' to refer to the collection of all information relevant to someone's interest as expressed by their query. The query here serves as one possible label to describe the general topic or an aspect of it. If we have some documents that were judged relevant to the query, we have a sample of information about the topic and – similarly to how language modeling uses samples – we can use that to estimate statistics of the complete topic. Lacking relevance judgments, we could use top-ranked documents retrieved in response to the query. These can then be further processed with a classifier for automatic labeling if required for the task at hand.

We note that because our experiments in Chapters 3, 4, and 5 use TREC corpora, it is impossible to avoid the use of topic as meant by TREC, even though we wish to use it differently. To make the distinction clear, when we refer to a 'TREC topic' we mean the written description of someone's interest, the sample queries, and usually the provided relevance judgments. When we use 'topic' elsewhere we are referring to the collection of information as described above.

2.2 Introduction to Diversity

A query may fail to retrieve relevant documents when its terms do not appear in any of the relevant documents. This is well-known in information retrieval (IR) as the 'vocabulary mismatch problem' (Girill, 1985; Furnas et al., 1987). Another related problem commonly occurs when the user expresses a query with a certain search intent in mind, but the formulation of the query may not be the best representation of that search intent. For example, the query may be ambiguous or underspecified. Then, the

user fails to retrieve relevant documents, and is often forced to reformulate the query. The reason for this problem is a knowledge gap (Verberne, 2011). The user does not know what keywords to choose to retrieve the information she is looking for. This happens since the user is unaware of what else she needs to know in order to acquire the desired knowledge. We need new strategies in IR for bridging this knowledge gap. Automatically reformulating a poorly represented query is one well-researched option (R. L. Santos et al., 2010a). Another option to alleviate this problem is presenting a more *diverse* set of search results to enrich the user's knowledge about the topic and to allow her to locate information beyond what she thought was relevant. This is applicable if for instance the query is underspecified rather than completely ill-formulated. Such diversified information will help the user to better reason about the *topic*. This leads us to *topical diversity*.

Topically diversifying search results means two things: (1) reducing topical redundancy among the results (C. L. Clarke et al., 2008); (2) presenting more varied relevant results to the user with each result putting emphasis on a different aspect of the topic. Both attributes complement one another. With result diversification the view of document relevance in IR is changed: in order to construct a result list of ranked documents for a given query, document relevance is calculated independently for each document according to the probability ranking principle (Robertson, 1977). That is, the relevance of each document to the query is estimated without taking other documents into consideration. However, in order to diversify or optimize a ranked list of search results, the relationships among documents must be taken into account. In this situation the whole ranked list of documents is one unit, whereas for traditional document ranking a single document is considered at a time. This new view expands the definition of document relevance and adds further factors to traditional ranking.

In addition to topical diversity, in certain situations users seek further qualities in results beyond relevance: particularly for debatable topics such as 'homosexuality', 'abortion', 'global warming', medical or political issues, users may be interested in factors such as opinionatedness or provocativeness as typically expressed in the questions 'What experiences did patients have with drug X?' and 'Are we responsible for global warming?'. For the latter question it might also be interesting for users to know what the general perception about global warming was ten years ago versus now, and how people's opinions changed over time. So this requires focus on the temporal aspect of search results. These are typical high-level aspects to be considered in search results, which are referred to as dimensions with non-topical aspects in this thesis. Note however that queries for which users look for non-topical qualities in the results do not have to be of debatable nature. They can also be strictly informative. For instance, 'tom tom gps' or 'central park' do not stem from highly provocative topics. Still, users may seek controversial information about these queries. Just as with topical diversification, we can rerank and improve results by favoring documents exhibiting the dimension in question. Instead of observing dimensions like opinionatedness and time in isolation though, considering them simultaneously can prove valuable: for instance, analyzing opinions at the time of a certain event versus other times when no significant events occur can yield interesting variations in the results. This leads us to "interestingness", which stands for the consideration of multiple dimensions with non-topical aspects. Below we briefly review prior works in these areas.

2.3 Topical Diversity and Other Related Work

How relevant is a document to a query if the user has already seen other retrieved documents on the topic? For this we need to know how novel the content of the document is. A large body of work related to topical diversity is that of novelty detection (Harman, 2002; Allan et al., 2003; Soboroff & Harman, 2005). The objectives

in the areas of topical diversity and novelty detection are very similar: reducing non-relevant and non-novel information from a retrieved list of documents or sentences.

Another very similar area to topical diversity and information filtering is Topic Detection and Tracking (TDT), in which topics are studied independently of the format of a ranked list for a given query (Allan, 2002). In the TREC Interactive Track (Hersh & Over, 1999), 'subtopics' are considered instead of a single, major topic that a document discusses. The objective is to cover the different aspects of relevance for a given topic. Evaluation measures used in the TREC Interactive Track are variations on set-based measures such as precision and recall, called 'aspectual precision' and 'aspectual recall' (Swan & Allan, 1998).

If a retrieved document is not novel, i.e., if it is too similar to a document that has been viewed before then this is redundant information for the user. Therefore, in existing literature topical result diversification refers to the elimination of topical redundancy in the results with the aim of maximizing the number of documents containing novel information (Carbonell & Goldstein, 1998; C. L. Clarke et al., 2008; Agrawal et al., 2009; R. L. Santos et al., 2010a, 2010b, 2011; Dang & Croft, 2012). By choosing a document that best exhibits relevance and novelty in each step, the reranked and diversified list is built *iteratively* in these prior works. This is a greedy approximation to the diversity problem, which was proved to be NP-hard with a reduction from maximum coverage (Agrawal et al., 2009). The approximation is within a factor of $1 - \frac{1}{e} \approx 0.63$ of the optimal solution. Achieving better than this approximation is known to be intractable unless P=NP (Feige, 1998).

How is (dis)similarity between documents computed? In one of the first works for topical diversification, Carbonell and Goldstein (1998) suggested a maximal marginal relevance (MMR) approach, which employs content-based similarity measures such as cosine similarity. Zhai et al. (2003) on the other hand proposed language modeling approaches such as mixture models and KL Divergence (Cover & Thomas, 1991). Other

methods in the literature are probabilistic (H. Chen & Karger, 2006) and correlation-based (Wang & Zhu, 2009). Very recent research addresses personalized diversification (Vallet & Castells, 2012), blog feed diversity (Keikha et al., 2012; R. L. T. Santos et al., 2012), and combined implicit and explicit aspect diversification (J. He et al., 2012).

More recently, researchers have shown explicit diversification approaches to be superior over implicit diversification techniques: well-known algorithms are xQuAD (R. L. Santos et al., 2010a, 2010b, 2011), IA-select (Agrawal et al., 2009), and more recently PM-1 and PM-2 (Dang & Croft, 2012). Among these approaches, it is common to equally or uniformly diversify across all query aspects or subtopics due to the lack of data (R. L. Santos et al., 2010a, 2010b, 2011; Dang & Croft, 2012). The proportionality models by Dang and Croft (2012) form an exception, since they are designed to work with any particular target distribution – whether uniform or non-uniform. In their experiments, due to the lack of suitable data, the authors were compelled to work with uniform distributions only. Although the TREC Web Track diversity task provides topical query aspects (C. L. A. Clarke et al., 2009), distributions over these aspects are not included. Agrawal et al. (2009) use their own classifiers and judgments for obtaining query intent aspect distributions. Further, the NTCIR-9 Intent task provides non-uniform aspect probabilities (Sakai & Joho, 2011). One of our contributions in this work is to present alternatives to the equal distribution approach (Section 3.2.3): a query's sentiment aspects distribution can be employed in various ways as a target bias to yield a certain emphasis in the results. Topical diversity could also benefit from these ideas.

Among the explicit diversification approaches, Agrawal et al. (2009) consider how well a document fulfills a certain query intent. Particularly ambiguous queries have multiple intents. In order to obtain a more varied results list, Agrawal et al. evaluate a document based on how well it addresses each query intent. These intents could also

be viewed as 'subtopics'. So topical diversity is achieved by ensuring that a maximum number of aspects or subtopics is covered in the reranked list, while documents heavily addressing an already included subtopic are penalized. This approach is more fine-grained than prior techniques, which assume that each document covers only one major topic. For this new diversification approach the authors also developed intent-aware versions of evaluation measures. Similarly, Carterette and Chandar (2009) proposed a probabilistic approach that considers the topics or relevance models of documents with respect to various query aspects.

Another fine-grained approach to diversification is that of R. L. Santos et al. (2010a), in which query reformulations are utilized instead of query intents for achieving more diverse search results. The authors present a new probabilistic framework called xQuAD for this purpose, in which the reranked list is built iteratively by choosing at each step the most relevant document with respect to the subqueries of the original query. Redundancy in the results is controlled by preferring highly relevant documents to subqueries that have not yet been addressed by other documents in the reranked list. We utilize the xQuAD framework with adaptations for our research described in Chapters 3, 4, and 5.

The proportionality-based approaches PM-1 and PM-2 (Dang & Croft, 2012) distinguish themselves from prior research by explicitly matching the aspect distribution in the diversified list to the overall popularity of these aspects, thus yielding a proportionally diversified list. Since these approaches have the capability to approximate a certain target bias, this forms a suitable basis for our work. We adapt this approach to sentiment and time diversity, and propose a variation for dealing with retrieval limitations (Section 3.2.2.3).

The notion of 'subtopic retrieval' has also been applied to topical diversity in Zhai et al.'s work (2003), in which set based measures are defined. These measures are modifications of precision and recall: *s-recall* and *s-precision*. They are more accu-

rate in evaluating coverage of various subtopics in a ranked list. The authors further introduce weighted subtopic precision, which is another version of s-precision incorporating a cost of redundancy. All these subtopic measures explicitly consider document ranks, which is a contrast to earlier set-based evaluation metrics *instance precision* and *instance recall* that were investigated in the TREC Interactive Track (Hersh et al., 2000). Other set-based evaluation metrics are presented at the idea level in Radlinski et al.'s work (2010). Finally, C. L. Clarke et al. (2008) present an adaptation of NDCG to the diversification framework: α-NDCG rewards novelty and diversity in addition to relevance.

2.4 Non-Topical Diversity

2.4.1 Opinionatedness

Our first dimension with non-topical aspects is opinionatedness. A document or topic can be regarded as opinionated or not depending on whether it contains subjective information. 'Opinionatedness' is a well-researched area within sentiment analysis and opinion detection in information retrieval (B. He et al., 2008; Zhang et al., 2007, 2008; Jia et al., 2009; Gerani et al., 2009; Seki & Uehara, 2009; Xu et al., 2011; R. L. T. Santos et al., 2012). One focus is on determining the polarity of the opinion in text; for instance whether a blog post tends towards the negative or positive side (Jia et al., 2009). Another emphasis is on how to achieve retrieval and opinion detection in a single step (Zhang et al., 2007; Gerani et al., 2009; Xu et al., 2011). The latter is referred to as 'opinion retrieval', which requires that search results not only be relevant, but also opinionated to a given query. A common approach to opinion retrieval is to start with retrieval, then the obtained documents are analyzed with respect to opinionatedness, and lastly, relevance and opinion scores are merged for reranking (Zhang et al., 2007; Gerani et al., 2009; Xu et al., 2011). Gerani et al. (2009) describe a learning to rank approach for combining relevance and opinion

scores, whereas Zhang et al. (2007) utilize several query document similarity signals for this combination step.

What these early prior works have in common is approaching opinion retrieval for blogs *without* considering diversity in search results. For evaluation, the given TREC Blog Track relevance judgments are employed as is with standard measures (Ounis et al., 2006; Macdonald et al., 2007). There is no evaluation that considers how well each result addresses a certain opinion, and how broad the spectrum of opinions in the results list is. Instead of a broad spectrum, the opinions may also be biased towards a certain direction. This is where opinion *diversity* comes into the picture. Research in this direction explicitly aiming at opinion diversification is very recent. Demartini and Siersdorfer (2010) describe a preliminary study about opinions in search results as given by popular search engines for controversial queries. They analyze how opinions are distributed in search results and arrive at the conclusion that search engines do not differ much in sentiment, but that higher ranked search results tend to be more positive than others.

In later work Demartini (2011) then tackles opinion diversification: his approach is based on the xQuAD framework (R. L. Santos et al., 2010a). Retrieved search results are classified into the sentiment categories positive, negative, or objective. For each query, there is a weighting for the importance of each of these sentiments. By using the xQuAD framework, search results are then reranked by choosing in each step the most topically and sentiment-wise relevant document. In order to avoid favoring the same sentiment, documents with similar sentiments to those already chosen into the reranked list are demoted. This way, documents with relevant, but yet *different* sentiments are preferred. We implement this approach as the SCS model (Chapter 3), as the TCS model (Chapter 4), and as DCS (Chapter 5) and combine it with different biases. The SCSF, TCSF and DCSF models presented in the same chapters are a further extension.

Kacimi and Gamper (2011) also propose an opinion diversification framework, in which they observe results for controversial queries, more specifically for informative and debatable queries in separation. As expected, the performance of their approach is slightly better for debatable queries. The authors consider three criteria for diversification: topical relevance, semantic diversification, and sentiment diversification. Their model favors documents most different in sentiment direction and in the arguments they discuss. The sentiments are again one of positive, negative, and neutral. In the model the components are linearly combined; however, in order to find the documents maximizing the distances for all criteria the authors exhaustively consider all subsets of documents, which takes exponential time. Our work differs from this work in several points: (1) We perform *sentiment diversification* only and not opinion diversification. Opinions refer to topical content, whereas sentiments are non-topical aspects that we focus on in this work. (2) This choice allows us to study sentiment diversification performance with different biases and different diversification frameworks, which has not been researched in prior work.

In this context, unlike topical diversity we make a simplifying assumption that each query represents one topical aspect. We avoid dealing with ambiguity by using long and specific queries in our experiments, as explained in Section 3.2.4. That is, the topical dimension is kept static so we can focus on the varied sentiment dimension. We leave it to future work to explore the interplay of topical and sentiment aspects together for diversification.

2.4.2 Time

Our second dimension with non-topical aspects is time. There has been a lot of work for integrating freshness and temporal aspects in general into information retrieval (Li & Croft, 2003; Sato et al., 2003, 2004; Uehara & Sato, 2005; Metzler et al., 2009; Dai & Davison, 2010; Dai et al., 2011). Sato et al. (2003) define fresh IR as

retrieving *current documents* having content from a certain time interval (Sato et al., 2003, 2004; Uehara & Sato, 2005). This includes documents written in the present about the present time. Temporal IR on the other hand refers to retrieving *any document* having content from a certain time interval. So typically such documents can be composed in the past or present about any time point or time range. Alonso and Gertz (2006) discuss how to use temporal information from documents for clustering. Kanhabua and Nørvåg (2010) on the other hand employ temporal information for reranking search results. One presented approach to this is through top-k retrieved documents for the input query using language modeling approaches. Kanhabua et al. (2012) also tackle determining the relevance of extracted times for an event. Sato et al. (2003) further define *fresh term frequency (FTF)*, which considers the freshness of terms as opposed to ordinary term frequency in IR. He presents three definitions and techniques for computing FTF for terms. These definitions also take into account how a document is modified over time, and which terms are added and deleted from it. Another, more static definition of FTF would be to assume that a document (such as a news article) is once published and does not change over time. Then, the terms in the document would only be associated with the publication or creation date of the document.

Pon et al. (2007, 2011) used freshness as a feature in their interestingness classifier. This feature measures the temporal distance between news articles by means of their publication dates. Other work deals with temporally organizing documents: Swan and Allan (2000) analyze terms for a specific time in order to obtain a timeline display of topics and events. 'Interesting' topics have a high χ^2 value, but relevance was not incorporated because the work was query-independent. Jones and Diaz (2007) also show how to construct temporal profiles of queries by examining the distribution of documents they retrieve. This is valuable for discovering interesting trends for a query and making informed decisions about which techniques to apply. The time dimension

has also been used in text summarization (Allan et al., 2001), where changes in news are observed over time so that only modified or new information is included in the generated summary.

Li and Croft (2003) define time-based language models for IR. They show how temporal information can be incorporated into the query likelihood and relevance models to improve performance for TREC queries. Freshness has also been used in web search (Dai & Davison, 2010) in two forms: a web page has its own freshness score which is decayed according to the activity on the page (e.g., creation or removal of links etc.), and it also has an in-link freshness score which is inferred from the activities of in-link web pages. Locating in-link web pages is similar to the PageRank algorithm (Brin & Page, 1998). The issue with PageRank is that it favors older pages because they typically have more in-links, so by considering times this factor can be dampened. Other works also use a variation of the PageRank algorithm with times (Yu et al., 2004). Very recent work includes learning to rank freshness and relevance in a framework for web search (Dai et al., 2011). The authors emphasize that freshness is more important for temporal queries which are about news or recent events, whereas it is less important or even harmful for time-insensitive or non-temporal queries. In their framework, temporal characteristics of a query are inferred and translated into features such that rankings are learned accordingly.

For blogs in particular, Keikha et al. (2011a, 2011b) showed query expansion techniques to be effective. These approaches employ terms from the most relevant days for the given query. Other work focuses on locating spiking time intervals for queries (Chasin, 2010; Kuzey & Weikum, 2012). We also use Wikipedia as a source for finding spiking time intervals for queries during experiments.

On the diversity front, Keikha et al. (2012) tackle temporal and topical diversity for blog feed retrieval. They compare blogs at the post level and penalize those with similar content or posting date. Not surprisingly, a combined approach using

topical and temporal diversity improves blog post retrieval. Similarly, Berberich and Bedathur (2013) present an approach to do temporal diversification on news documents: time aspects again come from document publication dates. In our work, temporal information comes from document content and not from posting or publication dates. This has several advantages: (1) relevant (pre-internet) times spanning a larger time period can be considered; (2) times are not necessarily 'points' representing only one relevant day, but rather time intervals, able to span time periods such as weeks, months, years, decades and beyond. Further, we investigate time (and sentiment) diversity not for retrieval effectiveness but for improved non-topical diversity, evaluating over sentiment aspects and time bins.

For temporal diversification we implicitly favor a certain type of temporal queries for our research: Jones and Diaz (2007) define *temporally ambiguous queries*, for which several time ranges and points are applicable. This stands in contrast to temporally unambiguous queries (only one time point or interval is relevant; this is often specified as part of a query, see Berberich et al. (2010)) and atemporal or non-temporal queries. Kulkarni et al. (2011) study the different types of changes and spikes in query profiles together with user search intents. For our work, studying temporal *non-topical* diversification makes most sense for temporally ambiguous queries since the results can flexibly be diversified across several time ranges. A typical user intent would be 'How did users think about this issue a couple of years (or months) ago versus now?' The time intervals can be varied here and possibly lead to interesting variety in opinionated search results. On the other hand, note that temporal topical diversification in the form of duplicate document detection would benefit all types of queries.

2.4.3 Interestingness

Finally, we review prior work for the topic "interestingness", which in this thesis serves as a general label for combining multiple dimensions with non-topical aspects. In Section 1.2.3 we instantiated interestingness with two dimensions – opinionatedness and time, but theoretically there can be others in addition or in place of these dimensions. Why do we use the label "interestingness" for this task? Interesting documents must somehow distinguish themselves from other ordinary documents about the topic to be able to catch a reader's attention. So 'interestingness' is a general term encapsulating many features that make the result 'interesting' or valuable without explicitly stating what these features are. This intuition is also backed up by the literature, in which interestingness has been studied in a user-dependent manner in Pon et al.'s (2007, 2011) work. A user-dependent interestingness classifier is trained for news articles. The findings from this work show that interestingness is a sufficiently complex characteristic that cannot be quantified in terms of a single feature or criterion: a combination of various features is required that should be weighted differently depending on the user in question. For this thesis, we are particularly interested in a *user-independent* definition of interestingness, and in the features playing a role in this. Perhaps the only related work in this direction is that of Allan et al. (2001): they define interestingness for news topics summaries by simply combining usefulness and novelty.

Measures for interestingness have been developed in the data mining community beyond information retrieval (Tan et al., 2002; Geng & Hamilton, 2006). The purpose in this prior work is slightly different than in ours, but still relevant: recognizing patterns or association rules that exhibit interestingness. Geng and Hamilton state that "so far there is no widespread agreement on a formal definition of interestingness", but that most important high-level aspects for characterizing interestingness of patterns are conciseness, coverage, reliability, peculiarity, diversity, novelty, surpris-

ingness, utility, and actionability (Geng & Hamilton, 2006). Some of these aspects are rather subjective, others are objective or content-based. The authors define these aspects and present a general framework for mining "interestingness" pattern rules. We use interestingness as a label to summarize several dimensions with non-topical aspects. In order to keep this research simple and clear, we analyze two dimensions only, but the list could certainly be extended to include more dimensions in further work.

Interestingness diversity – or the diversification of several dimensions with non-topical aspects – has not been researched yet in information retrieval to the best of our knowledge. The data mining community has explored diversity measures for interestingness to be applied to pattern discovery in databases but not to text (Hilderman & Hamilton, 2003). However, the aims are similar: reducing and reordering patterns presented to the user by eliminating redundancy in results. For this, the authors utilize statistical diversity measures and evaluate the distribution of the results for measuring variety.

With respect to combining sentiments and time in prior work, Tsytsarau et al. (2010) study the detection of opposing sentiments or contradictions over time, in particular with respect to scalability and representation of such fine-grained data. They show how to organize contradicting information in a time tree, and evaluate the usefulness of their automatic approaches against the human ability of finding contradictions in text. Our work studies sentiments and time for non-topical diversity: sentiments are not explicitly analyzed or filtered for contradictions. Since we use extreme biases for diversification, the obtained results emphasize varying kinds of sentiments in results over time – which can be anywhere in the spectrum between two contradicting extremes.

2.5 Summary

In this chapter we reviewed several areas of related work – topical diversity, novelty, and retrieval and diversity with various non-topical dimensions such as opinionatedness, time, and interestingness. Overall, it is evident that there is a good amount of research with non-topical dimensions for *retrieval*. However, this is not the case with *diversity*, having been mainly explored at the topical level and in limited amounts as to diversity. Our aim in this thesis is to fill this gap. Proposed explicit diversification approaches build on xQuAD (R. L. Santos et al., 2010a) and the Proportionality Model (Dang & Croft, 2012), considering both single non-topical dimensions for diversity, as well as their combination. Further, we adapt existing well-known evaluation measures from the diversity literature to work in this new setting.

CHAPTER 3

OPINIONATEDNESS

This chapter describes our work for the dimension opinionatedness. In order to more accurately capture the opinionatedness of a topic, we define a new measure, *average topic sentiment*. This measure is a combination of *provocativeness* and *balance*, which we first introduced in prior work (Cartright et al., 2009). After introducing the measures and showing a quick analysis on the TREC Blog Track (Ounis et al., 2006; Macdonald et al., 2007), we focus on how to achieve diversification for the sentiment dimension.

Part of this chapter has been published in our previous work (Aktolga & Allan, 2013). The diversification models and biases are formally introduced in Section 3.2.2. Then, we describe the experimental setup, supplementary tools, and the evaluation measures in Section 3.2.4 before presenting and discussing the results.

3.1 Measures at the Topic Level

In this section we define measures related to opinionatedness at the *topic level*. Below, first we introduce the various concepts involved, and Section 3.1.2 then uses this terminology for the measures.

3.1.1 Terminology

Query, Topic and Relevance In information retrieval, we refer to 'relevance data' or 'relevant units' to denote *truth data* with respect to a query or its topic T, henceforth abbreviated as $rel(T)$. Truth data is required as a gold standard for the evaluation of a system or approach. In this section, we assume that relevance data of some

form is available for the measures irrespective of which method was used to obtain it. Further, as detailed in Section 2.1, we assume that a query Q serves as a label or description for a topic T. Hence, when we refer to a 'topic', it is a query's topic $T(Q)$. For simplicity, we abbreviate $T(Q) = T$. For the topic 'abortion' for example, different queries could be generated such as 'pro life pro choice debate', 'terminating pregnancy', 'abortion' etc. We will focus on queries and their aspect distributions for diversification in Section 3.2.2 onwards.

With respect to opinionatedness, in this section a relevant unit of information to T is either subjective or neutral, depending on whether it exhibits a sentiment. A non-relevant unit of information is assumed to be neutral and therefore not subjective.

Sentiment For our measures at the *topic level*, we classify a relevant unit of information r further into four classes:

- $positive(r)$: if r exhibits positive sentiments;
- $negative(r)$: if r exhibits negative sentiments;
- $mixed(r)$: if the statements expressed in r are mixed, i.e., of positive and negative kind;
- $neutral(r)$: if r does not contain negative or positive statements.

From this we can define a function that determines the *sentiment* of a unit of information r as follows:

$$sent(r) = \begin{cases} -1 & \text{if } r \text{ is negative} \\ 1 & \text{if } r \text{ is positive} \\ 0 & \text{if } r \text{ is neutral or mixed} \end{cases} \tag{3.1}$$

Let us now apply these definitions to the set of relevant units $rel(T)$ of a topic T. According to the definitions above, we can decompose $rel(T)$ as follows:

$$rel(T) = P \cup N \cup M \cup O \tag{3.2}$$

where $P = \{r|r \text{ is positive}, r \in rel(T)\}$, $N = \{r|r \text{ is negative}, r \in rel(T)\}$, $M = \{r|r \text{ is mixed}, r \in rel(T)\}$, and $O = \{r|r \text{ is neutral}, r \in rel(T)\}$. Then, for determining the sentiment of T, we would typically average over the summed *sent* values:

$$sent(T) = \frac{\sum_{r \in rel(T)} sent(r)}{|rel(T)|} \tag{3.3}$$

In the experiments discussed in Section 3.1.3, we use existing opinionated truth data from the TREC Blog Track for the Opinion Finding task (Ounis et al., 2006; Macdonald et al., 2007) in order to estimate sentiment values.

Subjectivity A relevant unit that exhibits sentiments can be described as subjective. The counterpart to subjective is neutral: these are units of information not exhibiting opinions. We define the subjectivity of a unit of information r to be a binary value:

$$\text{subjectivity}(r) = \begin{cases} 0 & \text{if } r \in O \\ 1 & \text{otherwise} \end{cases} \tag{3.4}$$

From a theoretical point of view it may be more meaningful to define subjectivity as a real value in the range 0 to 1 to denote the strength of opinionatedness, but in practice it is very difficult to determine the degree of opinionatedness to this extent both manually and automatically. Therefore, we employ subjectivity as a binary value.

3.1.2 Measures

When a user is reading a document, it would be useful for her to know if it is about a highly subjective topic. For example, topics like "flag burning" and "NAFTA" have a high degree of subjective documents and the reader should proceed carefully. Such

a document or web page in isolation may appear completely reasonable, but often represents a biased perspective on the topic being discussed.

Based on the terminology defined in Section 3.1.1, we propose new metrics *provocativeness* and *balance* that suggest when the topic could be controversial (Cartright et al., 2009). These metrics are inspired by the use of 'provocativeness' and 'balance' in the vernacular. This means that a topic is interpreted to be provocative if there are many opinions on it. Similarly, a topic is considered to be 'balanced' if all positions or sentiments on it are equally represented. Inspired by these intuitive notions, **provocativeness** (PROV) of a topic measures the degree of subjectivity of the topic, which describes the quantity of subjective versus neutral content on the topic. Topics with a high provocativeness should caution a reader to seek multiple perspectives on the topic.

We formally define the PROV of T to be the proportion of subjective (versus objective) material on T. We then approximate it using all (relevant) units in T:

$$\mathbf{PROV}(T) = \frac{\sum_{r \in rel(T)} \text{subjectivity}(r)}{|rel(T)|} = \frac{|rel(T) \setminus O|}{|rel(T)|} \tag{3.5}$$

This measures the ratio of subjective units of information among all relevant units of information for a topic. A relevant unit of information can theoretically be a document, a paragraph or even single sentences. In this section, a 'relevant unit of information' is typically interpreted as a document, which is partly due to the dataset with which we test the measures in the next section.

The **balance** (BAL) of a topic T is the degree to which sentiments on the topic differ. We define it as:

$$\mathbf{BAL}(T) = \frac{(|P| + |M|) - (|N| + |M|)}{|P \cup N \cup M|} = \frac{|P| - |N|}{|rel(T) \setminus O|} \tag{3.6}$$

This measure describes the amount of imbalance between the negative and positive sentiments on a topic. Mixed sentiments are cancelled out since they count as both

positive and negative. Negative balance values indicate that the analyzed unit set contains more negative than positive content. Likewise for positive balance values. Note that by construction, the balance is bounded between -1.0 and 1.0. A value of 0 indicates that the topic is evenly balanced between units containing positive and negative sentiment. A reader that is aware of the calculated balance of a topic will be able to discern if a particular document is more or less likely to reflect the majority sentiment on a topic, if one exists. Again, we can interpret this measure as the difference in sentiment for the topic with respect to all subjective documents in the topic.

We would like to use these measures further to quantify the *sentiment of a topic*, i.e., the average sentiment of a set of documents about a particular subject matter. We can arrive at an average topic sentiment score by combining the measures provocativeness and balance. To recapitulate, the balance measure describes the *sentiment direction* of the topic, whereas provocativeness quantifies the *strength of subjectivity* of the topic. Although a topic may be extremely imbalanced to either direction, if its provocativeness is not very high, then only a few documents have extreme sentiments in this topic. The majority of the documents is not subjective, so in this case we would like to weaken the topic sentiment score accordingly. However, if a topic is highly provocative, the average topic sentiment should be closer to what the balance measure reflects. This is also consistent with the 'balanced case', i.e., when a topic is highly provocative but balanced (=0), then the average topic sentiment should be considered as 'mixed'. Basically, with the topic sentiment measure we want to observe the amount of imbalance between positive and negative sentiments in the topic *among all relevant documents* and not just among the subjective ones. This is so that the full size of the corpus is taken into account.

Therefore, we can define the *sentiment of a topic* as the expected average sentiment across all relevant units or documents:

$$
\begin{aligned}
TS(T) &= E[sent(T)] \\
&= \frac{1}{|rel(T)|} \cdot \sum_{r \in rel(T)} sent(r) \\
&= \frac{1}{|rel(T)|} \cdot (-1 \cdot (|N| + |M|) + 1 \cdot (|P| + |M|) + 0 \cdot |O|) \\
&= \frac{|P| - |N|}{|rel(T)|} \\
&= \frac{|rel(T) \setminus O|}{|rel(T)|} \cdot \frac{|P| - |N|}{|rel(T) \setminus O|} \\
&= PROV(T) \cdot BAL(T) \qquad (3.7)
\end{aligned}
$$

$TS(T)$ yields a score between [-1;1] and is thus comparable to a document's sentiment.

Note that our definition and approach to determining topic sentiment is different to that in some prior work (Hurst & Nigam, 2004; Eguchi & Lavrenko, 2006; Mei et al., 2007): the primary aim in these related works is to determine topical sentences, i.e., identifying characteristic sentences in a topics's documents that can then be used for algorithmically inferring the topic's overall sentiment. In contrast, here we use *labeled data* to quantify the measures provocativeness and balance. From this, the topic sentiment measure is inferred. So the correctness of our approaches depends on the quality of the labeling, which can be estimated in various ways, such as with relevance judgments or a sentiment classifier.

3.1.3 Analysis on Blog Track

We employ PROV, BAL and TS to characterize the subjectivity of the topics used in the TREC Blog Track for the Opinion Finding task (Ounis et al., 2006; Macdonald et al., 2007). We use the relevance judgments from the TREC 2008 Blog Track for calculating the measures for TREC topics 851 to 950 and 1001 to 1050. Each topic comes with a list of documents judged to be relevant to that topic, as well as whether

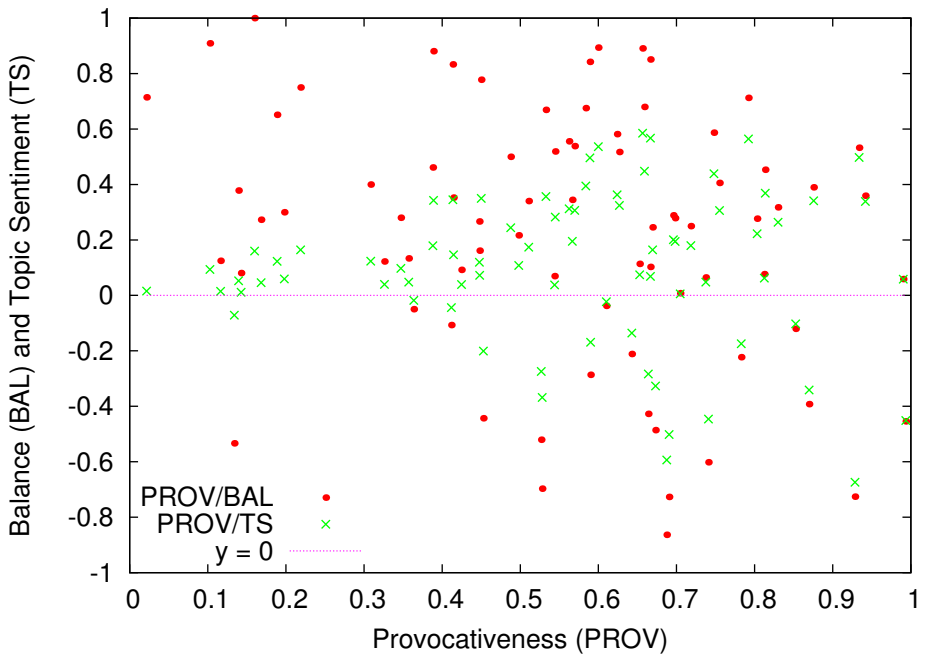

Figure 3.1. Provocativeness (PROV) against Balance (BAL) and Average Topic Sentiment (TS).

the assessor deemed the document positive, negative, mixed (containing both positive and negative content), or neutral (no sentiment). As defined by the data provided, a document acts as the unit of retrieval. A brief statistical summary of the topics is given in Table 3.1.

Table 3.1. Basic judgment statistics per TREC topic

Class:	*rel.*	*opin.*	*pos.*	*neg.*	*mixed*
min.	12	4	0	0	0
max.	950	826	392	533	455
avg.	292	182	70	56	57

Figure 3.1 shows a scatter plot of the three measures, PROV, BAL, and TS for the TREC blog topics. As expected, the topics used in the blog track tend to be provocative. We also observe that topics tend to express a higher degree of positive sentiments than negative. Further, TS is more skewed towards the x axis, dimming the balance factor according to the quantity of subjective content available on the

topic. Also note that for less provocative TREC topics, TS is much lower than BAL. Balance is more distributed towards the extremes 1 and -1.

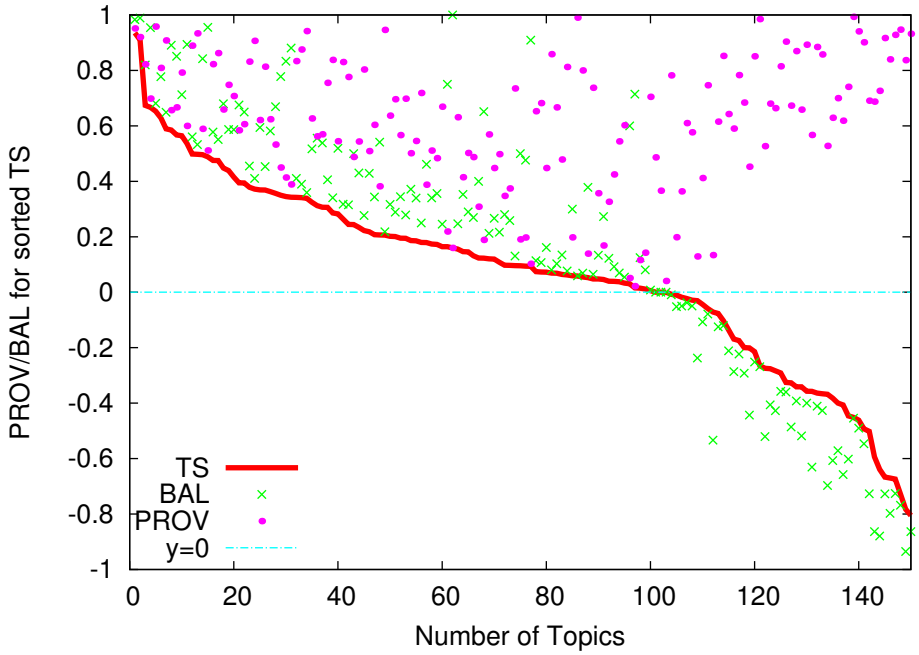

Figure 3.2. Provocativeness (PROV) and Balance (BAL) with decreasing Average Topic Sentiment (TS).

These findings become much clearer in Figure 3.2, where we observe provocativeness, balance and average topic sentiment for each of the 150 TREC topics (x axis), with the results sorted in decreasing order of TS. We can see that there is a large amount of positively opinionated topics, with most of them being mildly positive in the $TS = [0; 0.4]$ range. There is a slight, gradual decrease in provocativeness as TS gets lower, and provocativeness suddenly increases again for negative TS and BAL. In fact, for topics with negative sentiment and balance, most of them are moderately to highly provocative. When observing BAL together with the other measures, we can see that this is a very fluctuating measure – particularly for TREC topics with positive topic sentiment.

Table 3.2. TREC Topics with extreme and balanced topic sentiment (TS).

Topic	*Title*	*TS*	*PROV*	*BAL*
925	mashup camp	0.935	0.952	0.983
1021	Sheep and Wool Festival	0.910	0.920	0.989
864	colbert report	0.674	0.822	0.820
1012	Ed Norton	0.667	0.698	0.955
1032	I Walk the Line	0.653	0.958	0.681
855	abramoff bush	0.629	0.809	0.777
...	...	...	...	...
899	cholesterol	0.015	0.020	0.714
891	intel	0.015	0.116	0.125
862	blackberry	0.011	0.143	0.080
863	netflix	0.006	0.705	0.008
950	Hitachi Data Systems	0.000	0.040	0.000
927	oscar fashion	0.000	0.366	0.000
1005	Windows Vista	-0.007	0.782	-0.009
896	global warming	-0.010	0.199	-0.052
...	...	...	...	...
1017	Mahmoud Ahmadinejad	-0.639	0.726	-0.879
1013	Iceland European Union	-0.667	0.917	-0.727
1038	israeli government	-0.670	0.840	-0.798
870	“barry bonds”	-0.674	0.929	-0.726
867	cheney hunting	-0.728	0.947	-0.768
1031	Sew Fast Sew Easy	-0.783	0.837	-0.935
1008	UN Commission on Human Rights	-0.805	0.932	-0.864

The Average Topic Sentiment measure allows us to view the overall opinionatedness and direction of sentiment in one measure. Table 3.2 shows topics with extreme sentiment in either direction: highly negative topics rather touch debated political issues, which is not surprising. TREC topics with balanced sentiment tend to have lower provocativeness with the exception of ‘netflix’ and ‘Windows Vista’. In the next section, we move from the *topic-level* analysis to how these concepts apply in the context of *search and diversification*, where we typically have *queries* with their sentiment aspects distributions and documents with sentiments that are to be (re)ranked in response to the query.

3.2 Diversification

3.2.1 Introduction

In previous work diversification has mainly been applied for better topical variety in search results (Dang & Croft, 2012; R. L. Santos et al., 2010a, 2010b, 2011). Equal preference is typically given to all query aspects. How can opinionated content exhibiting sentiments be diversified? Initial approaches have been presented (Demartini, 2011; Kacimi & Gamper, 2011, 2012); however these only consider equal diversification across all query aspects. In this work, we view the problem from a high-level perspective to allow for sentiment diversification according to *different biases*, which will be vital for situations like a literature review on a controversial topic.

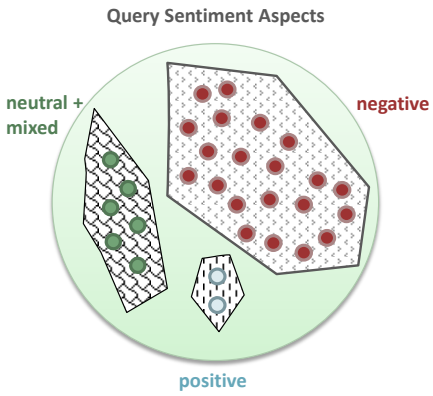

Figure 3.3. Query Sentiment Aspects: Dots represent units of information grouped by sentiments for this query: the obtained query sentiment aspects distribution is used for sentiment diversification.

Consider the query 'global warming.' In a typical use case, a user engages in a comprehensive literature review with the aim of understanding the positions on this topic. This involves – besides searching and finding relevant opinionated documents (Huang & Croft, 2009) – understanding and mentally categorizing opinionated content. This can be done by organizing the discussed arguments by topical content; or, they can also be grouped by sentiment, such as positive, negative, neutral, and mixed (Kacimi & Gamper, 2012). We focus on facilitating the latter approach for the

user. For some queries that can clearly be generalized into 'pro' versus 'con' arguments, this sentiment categorization is more natural, whereas it can be less obvious for queries like global warming that are associated with various arguments. Focusing on the *sentiment dimension* of these arguments, we can see that negative sentiments for global warming typically express criticism and concern about it and its effects on the environment. Those with positive sentiment often claim that worries about global climate change are unjustified ("there is no such issue"), playing down the concerns in a 'calming' (=positive) way. Mixed or neutral statements either express no sentiments or contain an equal amount of positive and negative arguments. Those could be "I don't care", or "It's a serious problem but we're handling it" kind of stances towards global warming.

Getting back to our use case: while a balanced and unbiased presentation of the results helps the user understand various viewpoints on a topic, discerning the topic's polarity is harder if minority opinions are 'buried' in the results (Kacimi & Gamper, 2012). Therefore, the user should be able to *switch the result perspective* as needed. This way, she can either obtain a balanced or a biased view on majority or minority opinions, make her own comparisons across the representations, and perform this task in a more informed manner. Note that this is different from showing all positive *or* all negative *or* all neutral/mixed documents at a time: with such a representation the user would still need to draw her own conclusions about which sentiments form majority or minority opinions. Our aim is to analyze this information for the user and to match the inferred trend as closely as possible in the results. For this, we need to have a good grasp of the query and its sentiment aspects distribution a priori: we analyze a large pool of data on the topic, which is grouped by sentiment aspects as visualized in Figure 3.3. Then, we can infer the *query's sentiment aspects distribution or inherent bias* from this analysis. We categorize the aspects 'mixed' and 'neutral' together to represent the 'balanced' aspect, whereas 'positive' and 'negative' refer to

arguments that are clearly biased towards one side only. If Figure 3.3 represented the Query Sentiment Aspects Distribution for global warming, this could be interpreted as the issue being perceived with great concern since negative sentiments constitute the majority, and while there are some 'balanced' positions on it, the positive sentiments form a clear minority. By utilizing this information during diversification, three target biases are emphasized in search results: (1) Equal diversification by preferring all sentiment aspects equally. This allows for a balanced representation of all sentiments; (2) Diversification towards the Query Sentiment Aspects Distribution, in which the resulting reranked list mirrors the estimated sentiment bias observed for the query's topic. This approach highlights the general perception of a topic; (3) Diversification against the Query Sentiment Aspects Distribution, in which documents about the minority sentiment(s) are boosted whereas those with the majority sentiment are demoted. Such a list highlights unusual and outlying opinions on the topic.

In this chapter we propose different diversification models for sentiment diversity with these 3 biases, and conduct experiments using the TREC Blog Track data (Ounis et al., 2006). Since sentiment classification is an essential tool for this task, we experiment by gradually reducing the accuracy of a perfect classifier down to 40%, and show which diversification approaches prove most stable in this setting. This shows the impact of sentiment classification on diversification performance. Further, in case the Query Sentiment Aspects Distribution cannot be reliably estimated (such as for a newly emerging topic, or when suitable data is not available (Dang & Croft, 2012)), we show how performance is affected by equal diversification when actually an emphasis either towards or against the Query Sentiment Aspects Distribution is desired.

3.2.2 Sentiment Diversification

3.2.2.1 Introduction

Given a query Q, we consider the countable query sentiment aspects $\sigma \in sent(Q)$, which are typically positive, negative, and neutral/mixed. We will use the distribution of sentiment aspects $\{\sigma_1, \ldots, \sigma_n\}$ for Q in our models to diversify search results accordingly. Sentiment aspects of the form positive, negative, and neutral/mixed can take different shapes when converted into a sentiment score for a document. In the literature (Demartini, 2011; Kacimi & Gamper, 2011; Ounis et al., 2006) we identified a document to either have a single discrete sentiment from $\{-1, 0, 1\}$, or the sentiment is broken down into three scores *positivity*, *negativity*, and *neutrality* such that they sum to 1.0 for a single document. We refer to these latter ones as finer grained "fractional scores" in the rest of the chapter. Our models are designed for these fractional scores, but discrete scores as introduced in Section 3.1.1 can also be handled by simple conversion as we will show later.

Below we consider two different diversification frameworks and present several modifications to them.

3.2.2.2 Retrieval-Interpolated Diversification

Algorithm 1 Retrieval Interpolated Diversification Framework.

$S = \emptyset$
while $|S| < \tau$ and $|R| > 0$
    do
        $D^* = \arg\max_{D \in R}\ \lambda \text{RetC}(Q) + (1 - \lambda)\text{SentC}(Q)$
        $R = R \setminus \{D^*\}$
        $S = S \cup \{D^*\}$
return S

Algorithm 1 shows the Retrieval-Interpolated Diversification Framework, which is similar to xQuAD, first introduced by R. L. Santos et al. (2010a) for topical diversity.

In this diversification framework, documents retrieved in R are iteratively added to the new ranked list S. The aim is to build a diversified list to boost relevant documents with important sentiments in the ranking without achieving a clustered representation of documents: documents emphasizing different sentiments should be distributed across the ranks to minimize redundancy. The τ documents are chosen according to the maximization objective function in line 4:

$$D^* = argmax_{D \in R} \; \lambda \cdot \text{RetC}(Q) + (1 - \lambda) \cdot \text{SentC}(Q) \tag{3.8}$$

where RetC(Q) is the *retrieval contribution*, which is always estimated with $P(D|Q)$ – how likely D is to be relevant to Q by content, and SentC(Q) is the *sentiment contribution*, which we will define in two different ways below. The scores from these two components are interpolated for diversity estimation.

3.2.2.2.1 Sentiment Contribution by Strength (SCS) In this version of the model we estimate the sentiment contribution in the maximization objective function (Equation 3.8) as follows:

$$\text{SentC}(Q) = P(D, \bar{S}|Q) \tag{3.9}$$

Here $P(D, \bar{S}|Q)$ measures how much D can contribute to the sentiment diversity of S. Structurally, this resembles xQuAD (R. L. Santos et al., 2010a).

In order to make the model more flexible towards sentiment scores, we define each document to have a fractional score for each sentiment aspect $\sigma \in sent(Q)$. For example, a document may be classified as positive with 75% confidence. Then, this can be converted into a trinary score $P(\sigma = \text{positive}|D) = 0.75$, $P(\sigma = \text{neutral}|D) = 0.25$, and $P(\sigma = \text{negative}|D) = 0$. Fractional classification scores directly obtained from a classifier (such as logistic regression) fit in nicely into this framework. If documents are manually judged, they are often associated with only one 'dominant' sentiment

score from $\{-1, 0, 1\}$ such as -1, which can be converted into a 100% negative score. Given this information, we can further decompose $P(D, \bar{S}|Q)$ as follows:

$$P(D, \bar{S}|Q) = \sum_{\sigma \in sent(Q)} P(D, \bar{S}|\sigma) \cdot P(\sigma|Q) \tag{3.10}$$

$$\stackrel{rank}{=} \sum_{\sigma \in sent(Q)} P(D|\sigma) \cdot P(\bar{S}|\sigma) \cdot P(\sigma|Q) \tag{3.11}$$

where $P(\bar{S}|\sigma)$ denotes the likelihood of σ not being satisfied by the documents already chosen into S (see below for further derivation) and $P(\sigma|Q)$ stands for the importance of sentiment aspect σ to query Q. This is discussed in detail in Section 3.2.3. From Equation 3.10 to Equation 3.11 we make the same independence assumption as R. L. Santos et al. (2010a): the diversity estimation of D with respect to the sentiments σ can be made independently of the documents already selected into S. Lastly, for practical purposes in the experiments we employ $P(D|\sigma) \stackrel{rank}{=} P(\sigma|D)$ by applying Bayes' Rule, and drop the constants. We continue with Equation 3.11:

$$\begin{aligned}
&\sum_{\sigma \in sent(Q)} P(D|\sigma) \cdot P(\bar{S}|\sigma) \cdot P(\sigma|Q) \\
&= \sum_{\sigma \in sent(Q)} P(D|\sigma) \cdot P(\sigma|Q) \cdot \prod_{D_j \in S} P(\overline{D_j}|\sigma) \\
&= \sum_{\sigma \in sent(Q)} P(D|\sigma) \cdot P(\sigma|Q) \cdot \prod_{D_j \in S} 1 - P(D_j|\sigma)
\end{aligned} \tag{3.12}$$

Here we make another independence assumption for $P(\overline{D_j}|\sigma)$ as R. L. Santos et al. (2010a): the likelihood of not sampling D_j's sentiment aspect from $sent(Q)$ is independent of the sentiments of the other documents in S. Since each D_j was independently chosen into S, this is a reasonable assumption.

To summarize, Equation 3.12 estimates the diversity of a document D by considering how well D represents each sentiment aspect, which is weighted by how important that sentiment aspect is to Q. This whole part is demoted according to how many documents of the same sentiment S already contains.

3.2.2.2.2 Sentiment Contribution by Strength and Frequency (SCSF) We consider an alternative formulation of the sentiment contribution component above in Equation 3.8 in which the punish/reward factor is estimated slightly differently:

$$\text{SentC}(Q) = P(D|Q) \cdot (1 - P(S|Q)) \tag{3.13}$$

where $P(D|Q)$ stands for how important D's sentiment is for Q, and $1 - P(S|Q)$ describes how well the sentiment aspects distribution for Q is already represented in S. We further derive:

$$\begin{aligned} & P(D|Q) \cdot (1 - P(S|Q)) \\ = \; & P(D|Q) - P(D|Q) \cdot P(S|Q) \\ = \; & \sum_{\sigma \in sent(Q)} P(D|\sigma) \cdot P(\sigma|Q) - P(D|\sigma) \cdot P(\sigma|Q) \cdot P(S|\sigma) \end{aligned} \tag{3.14}$$

Here we apply Bayes' Rule to $P(S|\sigma)$:

$$P(S|\sigma) = \frac{P(\sigma|S) \cdot P(S)}{P(\sigma)} \stackrel{rank}{=} P(\sigma|S) \tag{3.15}$$

which is rank-equivalent since $P(S)$ is a constant across all documents in an iteration, and $P(\sigma)$, the prior probability of a particular sentiment, is equal across all sentiments. Hence we obtain from Equation 3.14:

$$
\begin{aligned}
& \sum_{\sigma \in sent(Q)} P(D|\sigma) \cdot P(\sigma|Q) - P(D|\sigma) \cdot P(\sigma|Q) \cdot P(\sigma|S) \\
= & \sum_{\sigma \in sent(Q)} P(D|\sigma) \cdot P(\sigma|Q) \cdot (1 - P(\sigma|S)) \\
= & \sum_{\sigma \in sent(Q)} P(D|\sigma) \cdot P(\sigma|Q) \cdot P(\bar{\sigma}|S) \qquad (3.16)
\end{aligned}
$$

Now we can see that the first part of Equation 3.16 is identical to Equation 3.12. We can estimate the components $P(D|\sigma) \cdot P(\sigma|Q)$ the same way as described in Sections 3.2.2.2 and 3.2.3. However, $P(\bar{\sigma}|S)$, the likelihood of S not having sentiment aspect σ, is new. We define its complement as follows:

$$
P(\sigma|S) = \frac{sent(\sigma, S)}{|S|} \qquad (3.17)
$$

which is the number of documents in S having *dominant sentiment* σ. Each document in S can be mapped into its dominant or most confident sentiment class $\sigma \in sent(Q)$, typically positive, negative, or neutral/mixed. Given this, we count the number of times a particular sentiment σ occurs in S as $sent(\sigma, S)$. We set $P(\sigma|S) = 0$ if $S = \emptyset$ to avoid zero division in the first iteration.

To summarize, this formulation calculates the punish/ reward factor directly from the *frequency* of documents present in the whole set S with certain sentiments. Contrarily, in the SCS model the *strength of sentiments* of each document in S is considered individually, whereas the frequency of such documents is implicit in the multiplication over all documents in S. In the experiments we empirically verify the effectiveness of the two models in sentiment diversification to draw conclusions about their usefulness.

3.2.2.3 Diversity by Proportionality

Algorithm 2 Diversity by Proportionality (PM-2).

$S = \emptyset$
$\forall\sigma \;\; s_\sigma = 0$
while $|S| < \tau$ and $|R| > 0$
    do
        for $\sigma \in sent(Q)$
            do
                $quotient[\sigma] = \frac{v_\sigma}{2s_\sigma+1}$
        $\sigma^* = \arg\max_\sigma quotient[\sigma]$
        $D^* = \arg\max_{D \in R} \;\; \lambda \cdot quotient[\sigma^*] \cdot P(D|\sigma^*) + (1-\lambda) \sum_{\sigma \neq \sigma^*} quotient[\sigma] \cdot P(D|\sigma)$
        $R = R \setminus \{D^*\}$
        $S = S \cup \{D^*\}$
        for $\sigma \in sent(Q)$
            do
                $s_\sigma = s_\sigma + \frac{P(D^*|\sigma)}{\sum_{\gamma \in sent(Q)} P(D^*|\gamma)}$
return S

As a second diversification framework we consider PM-2 (Algorithm 2), the best-performing approach by Dang and Croft (2012). This framework is based on the Sante-Laguë method for seat allocation and is adapted here to sentiment diversification. The aim is to iteratively build the diversified list S from the retrieved list of documents R by maximizing the *proportionality* of S. The proportionality of S is tracked by means of two variables v_σ and s_σ, employed in Algorithm 2 in the quotient calculations. v_σ relates to the popularity of sentiment σ, denoting the number of relevant documents σ should have in S given a certain rank position. s_σ on the other hand represents the reality – the estimated number of documents *actually present* in S for σ given a certain rank position. We use the Query Sentiment Aspects information to estimate v_σ, which can easily be inferred from $P(\sigma|Q)$ at a particular rank i as follows:

$$v_\sigma = \lfloor i \cdot P(\sigma|Q) + 0.5 \rfloor \tag{3.18}$$

In the ideal S, the number of relevant documents for each aspect or sentiment is proportional to its popularity, represented by v_σ. This objective is similar to that of the Sante-Laguë seat allocation problem. Instead of applying the algorithm to topical aspects following Dang and Croft (2012), here it is employed together with sentiment aspects $\sigma \in sent(Q)$. We start with an empty S in line 1 in Algorithm 2. S has a maximum of τ seats available, which will be iteratively filled with documents. s_σ is initialized to 0 for all sentiment aspects σ in line 2. Then, for each of the τ seats, we compute the quotient for each sentiment σ according to the Sante-Laguë formula in lines 5-7. In line 8, the seat is assigned to the sentiment σ with the largest quotient. Then, in line 9, we choose the document that is most relevant to σ, while also bearing some relevance to other sentiments. The emphasis on these two components is controlled by means of the interpolation parameter λ. Note that $P(D|\sigma)$ is estimated by means of fractional sentiment scores as defined in Section 3.2.2.2 instead of estimating the relevance of the document with respect to a (sub)topical aspect. Under this modification from Dang and Croft's work (2012), a document is purely evaluated on the basis of its sentiments and not according to topical relevance. After choosing D^*, we then remove it from R and add it to S. In lines 12-14, the counts for s_σ are updated according to D^*'s fractional sentiment scores. This corresponds to each sentiment taking up a 'portion' of the seats in S. For the relationship between document sentiments and the Query Sentiment Aspects Distribution, please refer to Section 3.2.3.

3.2.2.3.1 Diversity by Proportionality with Minimum Available Votes (PM-2M) Unlike the seat allocation problem in a voting system, in a retrieved list of documents there is an additional constraining factor. The top K documents retrieved from a search system constitute the source for diversification, so it is possible that a particular sentiment is underrepresented in this list. Unless the system requests more documents, the desired proportionality in the diversified list may not

be optimally achieved with the current set of documents. In this situation, with respect to PM-2 the given votes v_σ overestimate l_σ, the *actual number of documents* with sentiment σ in the retrieved top K set. For a large enough rank K, this may result in a suboptimally diversified list where documents with an over-emphasized sentiment are exploited early in the ranks. Therefore, we propose a small change to the quotient defined in Algorithm 2:

$$quotient[\sigma] = \frac{\min(v_\sigma, l_\sigma)}{2s_\sigma + 1} \tag{3.19}$$

which ensures that the quotient does not over-emphasize the importance of a sentiment if data is missing in the retrieved list. This technique has a remote resemblance to disproportionate stratified sampling in that documents are chosen slightly differently than dictated by the Query Sentiment Aspects Distribution in favor of improved overall diversity. We refer to this modified diversification approach as PM-2M and compare its effectiveness to PM-2, SCS and SCSF in the experimental section 3.2.4.

3.2.3 Favoring Different Biases in Search Results

In the presentation of the diversification models above $P(\sigma|Q)$ plays a central role in defining which target sentiment bias is favored in search results. Intuitively, this component stands for the importance of sentiment σ to query Q. Below we present three different possible biases in search results that the estimation of $P(\sigma|Q)$ impacts.

3.2.3.1 Equal Sentiment Diversification (BAL)

This is our baseline approach, which does not give preference to any sentiment aspect, but weights them equally or uniformly. Therefore, this approach does *not* utilize information from the Query Sentiment Aspects Distribution. We set

$$P(\sigma|Q) = \frac{1}{|sent(Q)|} \tag{3.20}$$

which results in each sentiment criterion $\sigma \in sent(Q)$ to be considered equally important. We refer to this bias method as 'Balance' (BAL) in Section 3.2.4.

We assume that with this balanced estimation the SCS model is equivalent to Demartini's approach (2011). Since this detail is not explicitly described in their work, it is most reasonable to assume an equal bias as in prior research.

3.2.3.2 Diversifying Towards the Query Sentiment Aspects Distribution (CRD)

In this approach we choose to diversify the retrieved list *towards* the distribution of sentiments for the query. Such results strongly represent the crowd's opinion(s). For this, we need to obtain information about the distribution of sentiment aspects for Q from somewhere. We start with a set of documents that are related to Q, possibly from training data but more likely from a ranked list of documents believed relevant to the topic. Each of those is sentiment-tagged (manually or automatically) and mapped to its dominant or most confident sentiment class $\sigma \in sent(Q)$, so that we have $\{D_{pos}, D_{neg}, D_{neg}, D_{obj}, \ldots\}$. Given this, we count the number of times a particular sentiment σ occurs in Q's sentiment aspects distribution as $sent(\sigma, Q)$. This allows us to interpret $P(\sigma|Q)$ as the likelihood of sentiment σ being drawn from Q's sentiment aspects distribution:

$$P(\sigma|Q) = \frac{sent(\sigma, Q)}{\sum_{\varsigma \in sent(Q)} sent(\varsigma, Q)} \tag{3.21}$$

which represents the importance of sentiment σ to Q with respect to all sentiments ς. For instance, this is estimated through the fraction of positive (negative and neutral) documents observed for Q. We name this bias as 'Crowd' (short: CRD).

3.2.3.3 Diversifying Against the Query Sentiment Aspects Distribution (OTL)

What if a user is interested in viewing minority sentiments for the query? For favoring outlying sentiments, we need to diversify the search results *against* the Query Sentiment Aspects Distribution. For this, we introduce one minor modification to CRD above: Let the n sentiment estimations for $\sigma \in sent(Q)$ be sorted in increasing order of $P(\sigma|Q)$. Then, for each σ at rank i we swap its estimation $P(\sigma|Q)$ with the one at rank $n - i + 1$. This 'reverses' the values in the query sentiment aspects distribution without changing the properties of the distribution. Consequently, if originally in Q the positive sentiment is strongly favored and the negative sentiment is least favored, this trend is reversed through the value swap in the distribution so that outlying sentiments (negative sentiment aspect) will be strongly preferred during diversification. We refer to this bias as 'Outlier' (OTL) in the experiments (Section 3.2.4).

Irrespective of the preferred bias, we apply Add-1 Smoothing (S. F. Chen & Goodman, 1996) to $P(\sigma|Q)$ estimates to account for zero probabilities. In order to correct such unrealistic estimations, an unobserved sentiment class is assigned a very small probability, and the estimations for the other sentiment classes are adjusted accordingly.

3.2.4 Experiments

3.2.4.1 Setup

Retrieval Corpus As retrieval corpus we use the TREC Blog Track data from 2006 and 2008 (Ounis et al., 2006) for all our experiments. For preparation, the DiffPost algorithm is applied to the corpus for better retrieval as shown in prior work (Lee et al., 2008; Nam et al., 2009). Further, we perform stop word removal and Porter stemming.

Approach	P@10 (rel)	P@10 (op)	MAP (rel)	MAP (op)
QL title, $\mu = 2500$	0.604	0.438	0.310	0.230
QL description, $\mu = 2500$	0.618	0.423	0.259	0.193
SDM description , $\mu = 2500$	0.639	0.435	0.278	0.205
QL title & description, $\mu = 10000$	0.679	0.512	0.339	0.263
SDM title & description, $\mu = 10000$	**0.713**	**0.527**	0.373	0.285
SDM description, best passage, $\mu = 15000$	0.653	0.473	0.320	0.243
SDM title & description, best passage, $\mu = 15000$	0.705	0.521	0.377	0.288

Table 3.3. Retrieval experiments on the TREC Blog Track using all 150 queries.

Queries and Retrieval Model We split the 150 TREC Blog Track 2008 queries into 3 non-overlapping randomly chosen sets of size 50 each in order not to bias training or testing towards a specific year: split 1 is used for training and tuning parameters; the results in this chapter are reported on split 2, and split 3 is reserved for sentiment classifier training. For our diversification experiments, we use a strong retrieval baseline that we chose after some experimentation (Table 3.3): the queries' stopped title and description texts are combined for use with the Sequential Dependence Model in Lemur/ Indri (Metzler & Croft, 2005), smoothed using Dirichlet ($\mu = 10,000$). All diversification models are applied to the top $K = 50$ retrieved documents as determined during training. The retrieval scores are normalized to yield document likelihood scores.

Sentiment Classification The sentiment classifier is trained as a logistic regression model using Liblinear (Fan et al., 2008) with default settings, achieving 53.79% on split 1 and 49.25% on split 2. For training the model for three classes – positive, negative, and neutral – we utilize the judged documents from split 3. This yields probability estimates that are employed as fractional scores for sentiment estimation (Section 3.2.2.2.1). As features we extract Sentiwordnet 3.0 terms with their length-normalized term frequencies in the documents (Baccianella et al., 2010). We tried other training methods such as query-dependent feature selection versus query-independent feature selection; feature reduction; a different, small set of hand-crafted non-tf features; retaining only adjectives from the Sentiwordnet terms, and training a

Method	Accuracy on split 1	Accuracy on split 2
length-normalized tf with *all* Sentiwordnet terms, trinary	**53.79%**	**49.25%**
length-normalized tf with *all* Sentiwordnet terms, binary with threshold fitting for 3^{rd} class	50.83%	48.84%
length-normalized tf with *adjectives only* from Sentiwordnet, trinary, query-independent	51.45%	47.34%
length-normalized tf with *adjectives only* from Sentiwordnet, trinary, query-dependent	53.34%	46.50%
set of 11 non-tf features such as average pos./ neg./ neutrality scores, query-independent	53.32%	46.36%
set of 11 non-tf features such as average pos./ neg./ neutrality scores, query-dependent	49.35%	46.03%

Table 3.4. Some sentiment classifier accuracies on splits 1 and 2 with various training approaches.

trinary classifier versus a binary classifier with manual threshold fitting for the third class. All of these training methods achieve similar or worse performance, as shown in Table 3.4. Part of the challenge in training a successful sentiment classifier for this work is the lengthy nature of the documents in our corpus and the type of documents we apply the classifier to: blogs. A blog can contain a number of different sentiments of varying strengths: for example, completely different positions uttered by different users as comments, which makes the classification task of assigning one single score to a document harder – even for humans. Most sentiment classification research has been done on short texts like tweets or movie reviews (Pang, Lee, & Vaithyanathan, 2002; Turney, 2002). With sentence-long texts, the classification task is more straightforward since usually a single sentence contains one statement and therefore denotes a single sentiment position. With longer documents, apart from detecting sentiments across different sentences, the question of how to summarize those scores into one poses a hard challenge. Still, even single sentence classification

can become challenging in the presence of irony or sarcasm, for example (Davidov, Tsur, & Rappoport, 2010; González-Ibáñez, Muresan, & Wacholder, 2011). Standard commercial classifiers like SentiStrength[1] assign a neutral or mixed score of 0 to a sentence like 'Did you really think I'd let you do this?' or 'Did you really think I'd accept this?', which clearly hints at a negative sentiment of a person not being happy with a given situation. The person could be using such an utterance for scolding another person, for example. Therefore, this should clearly be classified as negative, and not neutral or mixed. Another misclassified example is the sentence 'I can't say I hate this.' which receives a negative score because of the negated auxiliary verb and the presence of the negative word 'hate'. Clearly, this sentence should be classified as mixed or positive because the person emphasizes that the object is not bad, i.e., negative. These examples clearly show that sentiment classification first has to be improved at the sentence level before attempting to improve performance at the document level.

Query Sentiment Aspects Distribution Estimation Given a query, its sentiment aspects distribution can be estimated in various ways: (1) in the form of opinion relevance judgments for a pool of documents where all judged relevant documents with respect to the query are included in the distribution. While this approach is very accurate for known queries, it cannot be applied to unseen queries; (2) by retrieving the top M documents from a separate corpus or web search engine with the query and tagging the documents with sentiment judgments. We experimented with both approaches: for (1) we use the relevance judgments from the TREC 2008 Blog Track (Ounis et al., 2006), which are divided into the same sentiment aspects as required in the models; for (2) we use the top 100 retrieved documents from a commercial search engine, which we tag with sentiments with our trained classifier.

[1] http://sentistrength.wlv.ac.uk/

3.2.4.2 Evaluation Measures

The sentiment diversification approaches are evaluated using standard evaluation measures that were designed for topical diversity: Precision-IA (Agrawal et al., 2009), s-recall (Zhai et al., 2003), α-NDCG (C. L. Clarke et al., 2008), ERR-IA (Ashkan & Clarke, 2011), and NRBP (C. L. Clarke et al., 2009). The former two measures are set-based, whereas the remaining ones are cascade measures as described by Ashkan and Clarke (2011), punishing redundancy through parameters α (α-NDCG, ERR-IA, NRBP) and additionally β (NRBP), which represents user patience. In order to measure sentiment diversity with a chosen bias, we implement all the measures in their intent-aware (or for us, 'sentiment-aware') version (Agrawal et al., 2009; Ashkan & Clarke, 2011). Hence, the weighted average over the sentiment-dependent scores of a measure is computed as given by measure-IA for a query Q:

$$\text{measure-IA}(Q) = \sum_{\sigma \in sent(Q)} P(\sigma|Q) \cdot \text{measure}(Q|\sigma) \tag{3.22}$$

where $P(\sigma|Q)$ defines the weight for the sentiment-specific result yielded by measure $(Q|\sigma)$.

Intent-aware measures can be rank-specific such as Precision-IA@k or α-NDCG@k for example, or rank-independent as NRBP. We utilize another rank-specific measure defined by Dang and Croft (2012), Cumulative Proportionality (CPR) at rank K:

$$CPR@K = \frac{1}{K} \sum_{i=1}^{K} PR@i \tag{3.23}$$

in which $PR@i$ is computed as the inverse normalized disproportionality at rank i:

$$PR@i = 1 - \frac{DP@i}{IdealDP@i} \tag{3.24}$$

Here, we define the disproportionality DP at rank i as follows:

$$DP@i = \sum_{\sigma \in sent(Q)} c_\sigma (v_\sigma - s_\sigma)^2 + \frac{1}{2} n_{NR}^2 \tag{3.25}$$

where v_σ is the number of relevant documents the sentiment σ should have at rank i, s_σ is the number of relevant documents actually found for σ until rank i, n_{NR} is the number of documents that are non-relevant (to any sentiment) until rank i, and $c_\sigma = 1$ if $v_\sigma \geq s_\sigma$, 0 otherwise. This measure allows us to assess how proportional the diversified list is with respect to the desired query sentiment aspects distribution. v_σ can be inferred from the true query sentiment aspect distribution $P(\sigma|Q)$ in the same way as detailed in Equation 3.18. As noted by Dang and Croft (2012), CPR penalizes the under-representation of aspects (here: sentiments) and the over-representation of non-relevant documents.

3.2.4.3 Results

In this section we discuss the results of the retrieval baseline SDM and all the diversification models in Section 3.2.2, SCS, SCSF, PM-2 and PM-2M, with the three target biases, Crowd (CRD), Balance (BAL) and Outlier (OTL). The interpolation parameter $\lambda \in \{0.0, ..., 1.0\}$ is tuned in 0.1 steps separately for each model and bias on our training split. The results are presented with fixed parameters K and λ on test split 2, and the evaluation is performed with the TREC 2008 Blog Track judgments at rank 20. α-NDCG, ERR-IA, and NRBP require parameters, which are set to $\alpha = 0.5$ and $\beta = 0.5$.

3.2.4.3.1 Straight-Bias Experiments Our primary aim in the experiments is to evaluate sentiment *diversification* performance. Sentiment classification is an important part of the system since both the to-be-diversified documents need to be tagged with sentiments, as well as those for the topic sentiment distribution estimation. Since a 'full evaluation' of sentiment diversification techniques on a publicly available dataset has not been done yet in prior work, it is important to understand

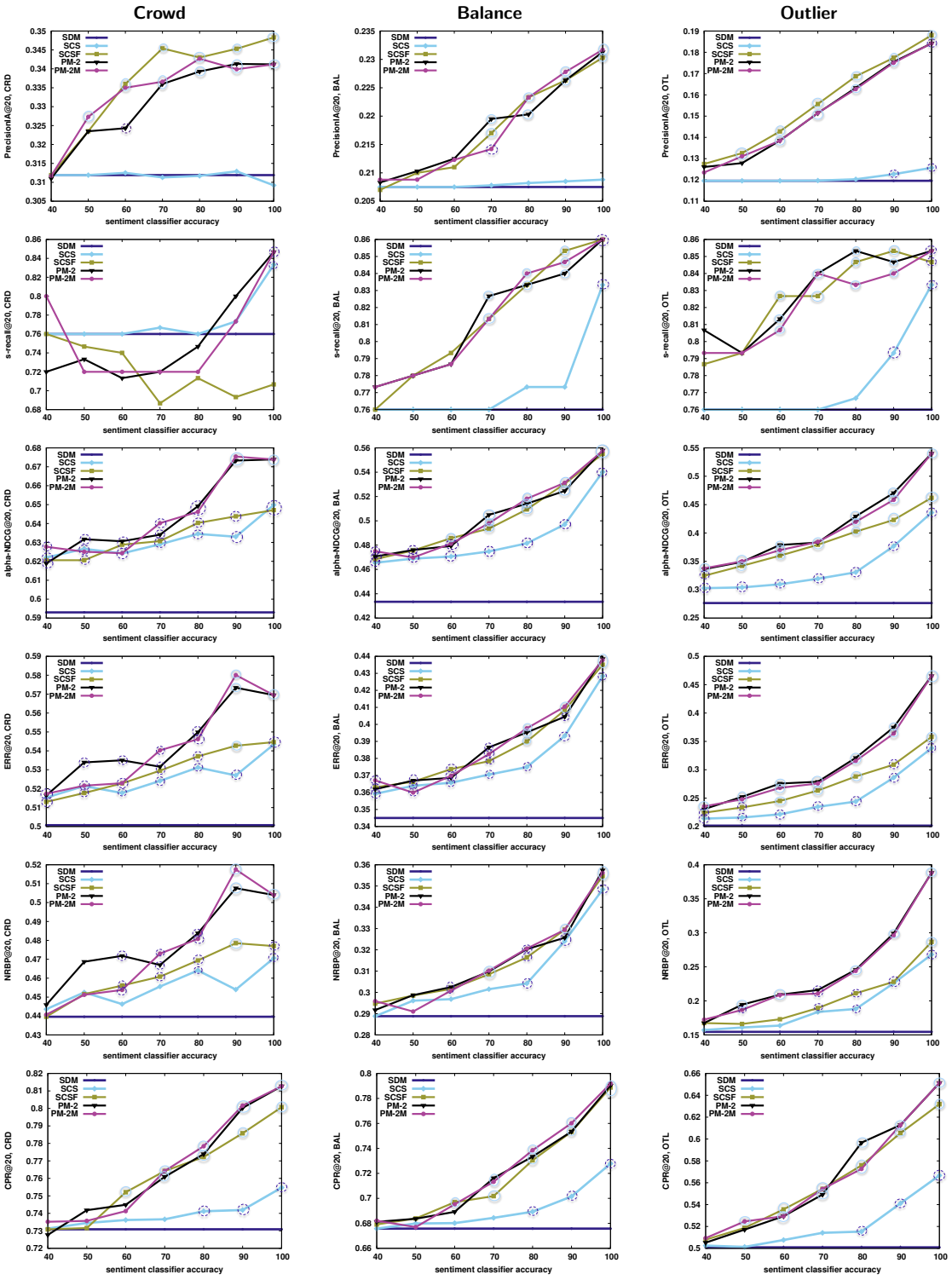

Figure 3.4. Straight-Bias Experiment over test split varying sentiment classifier accuracies on the x-axis and each one measure and bias on the y-axis. The leftmost column is for the Crowd bias (CRD), the middle one for Balance (BAL), and the rightmost one for Outlier (OTL).

how sentiment classification quality affects diversification performance. Therefore, we start with a "perfect system" in which classification accuracy is 100% for judged documents. For unjudged documents the trained sentiment classifier described in Section 3.2.4.1 is applied. We then gradually reduce the overall classification performance in 10% steps until 40% as follows: given the top $K = 50$ retrieved documents for a query, before diversification we randomly sample the ranks at which the true classification label is switched to another label randomly to achieve the desired classification error for each query.

Figure 3.4 shows the results for the straight-bias experiment, in which the query sentiment aspects distribution employed in experiment and evaluation *underlies the same favored bias*. For instance, the left-most column in Figure 3.4 shows the results for diversifying towards Crowd in the experiments, and measuring performance for Crowd in the evaluation (short: CRD-CRD). The middle column shows the same for Balance (short: BAL-BAL), and the right-most column is for the Outlier bias (short: OTL-OTL). Below, we show bigger versions of some of these graphs that are discussed in more detail.

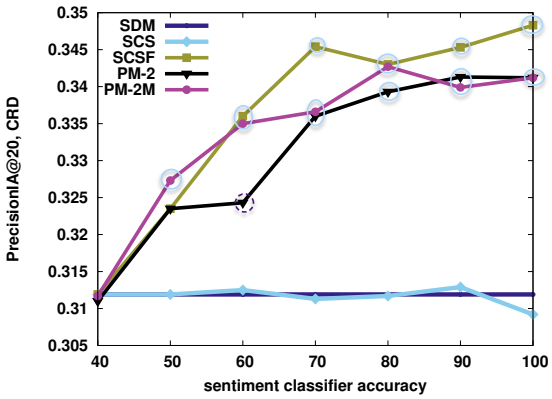

Figure 3.5. Precision-IA@20 results for the Crowd bias.

At the top-most row in the Precision-IA@20 graphs we observe a big gap between the SDM baseline and SCS model versus the rest of the models. For Crowd, the SCSF

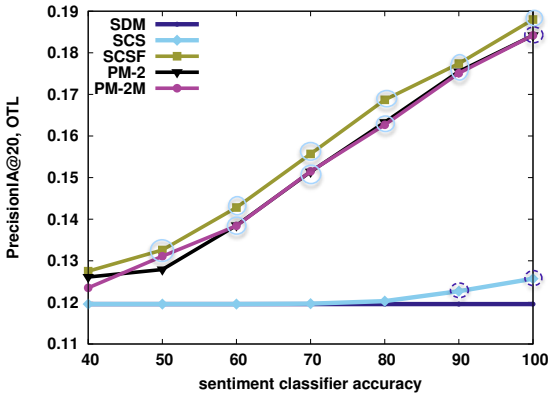

Figure 3.6. Precision-IA@20 results for the Outlier bias.

model only dominates when classification accuracy is at least 60% while it achieves the best (however not statistically significant) numbers in the Outlier graph (also see Figures 3.5 and 3.6). PM-2M and PM-2 also perform well and dominate some of the lower accuracy ranges. Statistical significance with the paired two-sided t-test (p-value < 0.05) is indicated in the graphs with circles: the lighter blue circles refer to the result being significant over the SCS and SDM models, whereas the darker dotted circles indicate significance over the SDM model only. In the Precision-IA@20 graphs the results for SCSF and the proportionality-based methods are significant over SCS and SDM even for lower accuracies. We conclude that if precision is important, the SCSF diversification model should be used.

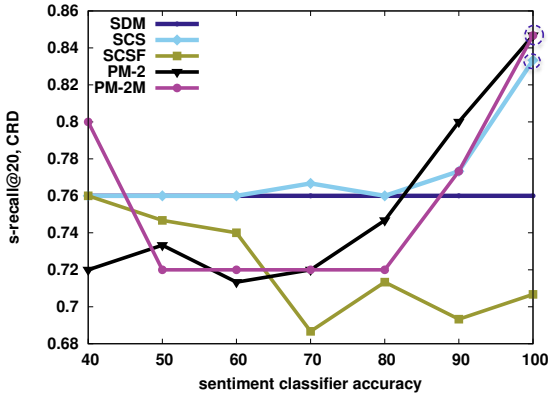

Figure 3.7. s-recall@20 results for the Crowd bias.

Among the s-recall@20 graphs the one for Crowd is the most arbitrary one (Figure 3.7). Performance drops well below the baseline for the SCSF and proportionality-based methods with medium quality classification: this indicates that the majority sentiment(s) are being emphasized too strongly, whereas minority sentiments appear much later in the ranked list for the first time, which is when the subtopic-recall measure is affected. This is expected, since we explicitly diversify in favor of majority sentiments. In the Balance and Outlier graphs for s-recall@20 there is no such trend, however precision is not as high for those biases as it is for Crowd. This is a typical precision versus recall tradeoff observation.

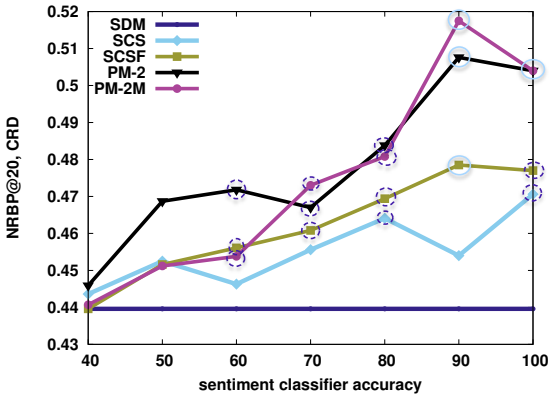

Figure 3.8. NRBP@20 results for the Crowd bias.

The next row shows results for α-NDCG@20, followed by ERR-IA@20 and NRBP: we note that the trends in these graphs look very similar, although the ranges of the values differ greatly. It is interesting to observe that the peak performance for the proportionality-based methods for the Crowd bias is not at 100% classification accuracy, but at 90% (Figure 3.8). What these three measures have in common is punishing redundancy based on the rank and sentiment criterion in addition to non-relevance. Since usually there are many documents with the majority sentiment in the retrieved list to start with, a strong emphasis on a single sentiment criterion results in more redundancy. With the 10% error in classification documents with other

sentiments are slightly boosted, yielding better overall varied ranking. In the Balance and Outlier graphs this trend cannot be observed, since the Balance bias does not strongly emphasize a single sentiment criterion to begin with. Concerning the Outlier bias, there are fewer documents with minority sentiments in the retrieved list to cause the same 'clustered' ranking effect as for Crowd. Summarizing the trends across the α-NDCG@20, ERR-IA@20, and NRBP graphs we make the following conclusion: if ranking is important, the PM-2 and PM-2M methods should be chosen.

Finally, we look at the last row of graphs with the CPR@20 results: this measure evaluates how proportional the overall list is with respect to the chosen bias. PM-2 and PM-2M achieve the best results, which is closely followed by SCSF. PM-2 and SCSF are more appropriate for lower classification accuracies ($\leq 70\%$), whereas PM-2M performs slightly better with better classification quality.

Method	Bias	λ, 100%	λ, 90%	λ, 80%	λ, 70%	λ, 60%	λ, 50%	λ, 40%
SCS	CRD	0.9	0.3	0.8	0.9	0.9	0.8	0.9
SCSF	CRD	0.6	0.5	0.3	0.2	0.9	0.3	1.0
PM-2	CRD	0.9	0.8	0.8	0.8	0.9	0.8	0.8
PM-2M	CRD	0.9	1.0	1.0	0.9	0.4	0.9	0.4
SCS	BAL	0.8	0.8	0.1	0.7	0.8	0.8	1.0
SCSF	BAL	0.5	0.6	0.7	0.8	0.3	0.0	0.8
PM-2	BAL	0.9	0.8	0.9	0.9	0.7	0.6	0.5
PM-2M	BAL	0.9	0.9	0.8	0.9	0.7	0.3	1.0
SCS	OTL	0.5	0.2	0.3	0.9	0.6	0.8	0.9
SCSF	OTL	0.1	0.1	0.3	0.8	0.0	0.9	0.9
PM-2	OTL	0.9	0.6	0.9	0.5	0.3	0.8	0.8
PM-2M	OTL	0.9	0.9	0.8	0.8	0.9	0.5	0.8

Table 3.5. Fixed λ values chosen for each method, bias and classifier accuracy.

Looking at the fixed values of the interpolation parameter λ during training in Table 3.5, the following insights can be drawn: for the SCS model, across all classifier accuracies and biases generally $\lambda \geq 0.6$ values are preferred. So this model performs best with a weaker emphasis on diversity, which pulls it closer to the SDM baseline as observed in the graphs of Figure 3.4. SCSF on the other hand has a good mixture

of higher and lower λ values across classifier accuracies and biases, with many of them being < 0.5, particularly when the classifier is more accurate. So a heavier emphasis on the diversification part helps this model. The distinguishing feature between SCS and SCSF is the consideration of sentiment frequencies in addition to sentiment strength contributions. When the classifier is noisy however ($< 60\%$) and thus sentiment frequency counts are not accurate, SCSF also benefits from higher λ values. In the PM-2 and PM-2M models the role of λ is different: it balances the emphasis on the chosen aspect $\sigma*$ versus all the other aspects $\sigma \in sent(Q), \sigma \neq \sigma*$. Here, consistently higher λ values are preferred for both models, i.e., a high emphasis on the chosen aspect and a minimal weight on the other ones seems most beneficial. The effectiveness of these two models solely relies on sentiment estimations: given our adaption of PM-2 from its original definition (Dang & Croft, 2012) to sentiment diversity, the retrieval scores are not used for building the diversified list.

Measure	Bias	Precision-IA@20	s-recall@20	α-NDCG@20	ERR-IA@20	CPR@20	NRBP
SDM	CRD	0.312	0.760	0.593	0.501	0.731	0.440
SCS	CRD	0.312	0.760	**0.622**	**0.515**	0.730	0.444
SCSF	CRD	0.294	0.767	0.606	0.493	0.694	0.415
PM-2	CRD	0.325	0.780	**0.624**	0.517	0.750	0.443
PM-2M	CRD	0.324	0.773	**0.618**	0.512	0.747	0.441
SDM	BAL	0.208	0.760	0.433	0.345	0.676	0.289
SCS	BAL	0.207	0.753	**0.457**	0.350	0.666	0.281
SCSF	BAL	0.209	0.793	**0.472**	**0.360**	0.676	0.289
PM-2	BAL	0.200	0.787	**0.467**	0.355	0.663	0.283
PM-2M	BAL	0.204	0.800	**0.466**	0.350	0.658	0.275
SDM	OTL	0.120	0.760	0.277	0.202	0.501	0.155
SCS	OTL	0.122	0.753	**0.298**	0.208	0.492	0.152
SCSF	OTL	0.121	0.780	**0.312**	0.210	0.485	0.145
PM-2	OTL	0.117	0.760	0.302	0.204	0.471	0.147
PM-2M	OTL	0.121	0.767	**0.319**	0.215	0.496	0.154

Table 3.6. Straight-Bias Experiments with trained classifier for all the models with different biases.

3.2.4.3.2 Straight-Bias Experiments with Trained Classifier In this experiment we use our trained sentiment classifier for tagging retrieved documents as well as for estimating the Query Sentiment Aspects Distribution. For the latter, the top 100 documents are retrieved from a commercial search engine with the TREC Blog

Track title queries. Diversification with all models and 3 biases yields the results presented in Table 3.6. Given the low quality of the classifier as shown in Table 3.4, here we can observe minor improvements over the SDM baseline, however very few of the results are statistically significant. Statistically significant results over the SDM baseline are printed in bold font (p-value < 0.05): only α-NDCG@20 and ERR-IA@20 are significant for some of the models. These results prove that having a high-quality sentiment classifier is crucial for sentiment diversification, since it affects the calculation of the bias and the classification of retrieved documents.

Measure	**Precision-IA@20**		**α-NDCG@20**		**s-recall@20**	
Exp-Eval	BAL-CRD	CRD-CRD	BAL-CRD	CRD-CRD	BAL-CRD	CRD-CRD
SDM baseline	0.312	0.312	0.593	0.593	0.760	0.760
SCS	0.308	0.309	0.642	**0.650**	0.833	0.833
SCSF	0.298	**0.348**	0.648	0.647	0.860	0.707
PM-2	0.302	**0.341**	0.642	**0.674**	0.860	0.847
PM-2M	0.298	**0.341**	0.639	**0.674**	0.860	0.847
Measure	**ERR-IA@20**		**CPR@20**		**NRBP**	
Exp-Eval	BAL-CRD	CRD-CRD	BAL-CRD	CRD-CRD	BAL-CRD	CRD-CRD
SDM baseline	0.501	0.501	0.731	0.731	0.440	0.440
SCS	0.532	**0.543**	0.750	**0.755**	0.453	**0.471**
SCSF	0.533	**0.545**	0.774	**0.801**	0.456	**0.477**
PM-2	0.526	**0.570**	0.772	**0.813**	0.446	**0.504**
PM-2M	0.521	**0.570**	0.767	**0.813**	0.440	**0.504**

Table 3.7. Cross-Bias Experiment over test split with perfect sentiment classifier to compare performance loss when diversifying equally (BAL-CRD) if actually diversification for the Crowd bias is desired (CRD-CRD).

3.2.4.3.3 Cross-Bias Experiments Consider the following real-world setting: for certain queries, it may not be feasible to collect data for estimating the distribution across query sentiment aspects, or suitable corpora may currently not be available. This could happen if the query represents a very recent event or topic and the data is not substantial enough for drawing general conclusions. As we have seen in this chapter, current state-of-the-art sentiment classifiers do not perform well, so

distributions may not be accurately estimated. In such a situation we can fall back to the Balance bias or equal diversification approach (Dang & Croft, 2012; R. L. Santos et al., 2010a, 2010b, 2011). Naturally, the next question to answer is how much performance is lost when diversifying with Balance instead of the desired bias such as Outlier. The cross-bias experiments in this section investigate this case, and enable us to draw conclusions about the value of collecting and using information about topic sentiment distributions for controversial topics.

We analyze two cases. The first, presented in Table 3.7 shows the results for equally diversifying for Balance, but performance is measured for the Crowd bias (BAL-CRD). This is contrasted with diversifying for the Crowd bias, and evaluating for the same (CRD-CRD). Bold entries in CRD-CRD indicate statistical significance over BAL-CRD with a p-value of < 0.004 (t-test, as before). The SDM baseline is included for comparison. As observed earlier, s-recall@20 results slightly decrease. All other CRD-CRD results for the proportionality-based methods are significant over BAL-CRD results, whereas for the SCSF and SCS models there are a few exceptions. We observe a maximum loss of 16.92% for Precision-IA@20 with SCSF, and an average loss of 6.48% across all measures and diversification approaches.

The second case is presented in Table 3.8: we observe the results for equally diversifying for Balance, but performance is measured for the Outlier bias (BAL-OTL). This is contrasted with diversifying for the Outlier bias, and evaluating for the same (OTL-OTL). Similar to Table 3.7 the results are statistically significant with p-value < 0.05 for OTL-OTL over BAL-OTL, but the losses with equal diversification are more heavily pronounced here: there is a maximum loss of 48.79% for NRBP with PM-2M, and an average loss of 16.23% across all measures and diversification approaches. So for highlighting minority sentiments through diversification it is even more important to be able to accurately predict query sentiments aspect distributions than it is for emphasizing majority sentiments as observed in Table 3.7. This

Measure	**Precision-IA@20**		**α-NDCG@20**		**s-recall@20**	
Exp-Eval	BAL-OTL	OTL-OTL	BAL-OTL	OTL-OTL	BAL-OTL	OTL-OTL
SDM baseline	0.120	0.120	0.277	0.277	0.760	0.760
SCS	0.126	0.126	0.413	**0.436**	0.833	0.833
SCSF	0.164	**0.188**	0.433	**0.462**	0.860	0.847
PM-2	0.166	**0.184**	0.447	**0.540**	0.860	0.853
PM-2M	0.166	**0.184**	0.446	**0.540**	0.860	0.853
Measure	**ERR-IA@20**		**CPR@20**		**NRBP**	
Exp-Eval	BAL-OTL	OTL-OTL	BAL-OTL	OTL-OTL	BAL-OTL	OTL-OTL
SDM baseline	0.202	0.202	0.501	0.501	0.155	0.155
SCS	0.309	**0.338**	0.562	**0.567**	0.237	**0.268**
SCSF	0.320	**0.358**	0.624	0.632	0.243	**0.287**
PM-2	0.337	**0.465**	0.632	**0.651**	0.262	**0.388**
PM-2M	0.336	**0.465**	0.634	**0.651**	0.261	**0.388**

Table 3.8. Cross-Bias Experiment over test split with perfect sentiment classifier to compare performance loss when diversifying equally (BAL-OTL) if actually diversification for the Outlier bias is desired (OTL-OTL).

way diversification can be performed with the intended bias rather than with equal diversification, which yields significantly worse results.

We presented the cross-bias experiments with perfect sentiment classification to reveal the maximum performance loss. As classification accuracy degrades, the losses become smaller but remain noticeable.

3.2.4.3.4 Analysis with Specific Queries To see the models in action, we look at the output for one query in Tables 3.9 and 3.10, number 1007 from the TREC Blog Track: 'women in Saudi Arabia', asking for opinions about the treatment of women in Saudi Arabia. We show titles or characteristic excerpts from the documents together with their overall sentiment. The query has the following sentiment aspects distribution: 67% negative, 17% mixed/neutral, and 16% positive. Here we diversify for the Crowd Bias, so the aim is to mirror this distribution in the results. The top 10 retrieved results with the SDM baseline are presented at the top left: this result list does not include any positive documents, and an equal amount of

	SDM baseline	
Rank	Excerpt	Sent.
1	The Religious Policeman: Mutt the Muttawa	-
2	Happy Feminist: PROTESTING GENDER...	o
3	Between tradition and demands for change	o
4	Saudi mobile carriers ban SMS voting...	-
5	Saudi Arabia, Ever Our Friends And Allies	-
6	Orientalism and Islamophobia	o
7	Laws discriminate against women...	-
8	...who urged SA to improve women's rights...	o
9	Being a Child in Saudi Arabia	o
10	Depressing Post: ...woman filed a case against...	-
	SCS	
Rank	Excerpt	Sent.
1	The Religious Policeman: Mutt the Muttawa	-
2	Happy Feminist: PROTESTING GENDER...	o
3	First women to win in Saudi elections	+
4	Between tradition and demands for change	o
5	Saudi mobile carriers ban SMS voting...	-
6	Saudi Arabia, Ever Our Friends And Allies	-
7	Orientalism and Islamophobia	o
8	Laws discriminate against women...	-
9	...who urged SA to improve women's rights...	o
10	Being a Child in Saudi Arabia	o

Table 3.9. Crowd Bias: Top 10 results with SDM baseline and SCS model for query number 1007, 'women in Saudi Arabia.' - denotes a negative document, o refers to mixed/neutral, and + to positive.

negative and mixed/neutral documents, which is clearly unsatisfactory for a Crowd bias representation of the results. The SCS model includes one positive document at rank 3, since lower ranked documents through the SDM baseline can be pulled up by the diversification models. Although the documents are nicely shuffled around across ranks, the ratio of the sentiments is still not close to the desired target query sentiment aspects distribution. The SCSF model is able to correct this, explicitly considering the frequency of documents with their dominant sentiments: we have 6 negative documents, 3 mixed/neutral, and 1 positive. But 4 negative documents are clustered right after each other, which slightly affects measures such as α-NDCG@10.

	SCSF	
Rank	Excerpt	Sent.
1	The Religious Policeman: Mutt the Muttawa	-
2	Happy Feminist: PROTESTING GENDER...	o
3	Saudi Arabia, Ever Our Friends And Allies	-
4	Orientalism and Islamophobia	o
5	First women to win in Saudi elections	+
6	Laws discriminate against women...	-
7	Depressing Post: ...woman filed a case against...	-
8	Their shabby treatment of women...	-
9	Oprah is being smuggled into Saudi Arabia...	-
10	Between tradition and demands for change	o

	PM-2	
Rank	Excerpt	Sent.
1	The Religious Policeman: Mutt the Muttawa	-
2	Saudi Arabia, Ever Our Friends And Allies	-
3	First women to win in Saudi elections	+
4	Happy Feminist: PROTESTING GENDER...	o
5	Orientalism and Islamophobia	o
6	Laws discriminate against women...	-
7	Depressing Post: ...woman filed a case against...	-
8	Their shabby treatment of women...	-
9	Thumps up for the Saudi ladies.	+
10	Between tradition and demands for change	o

Table 3.10. Crowd Bias: Top 10 results with SCSF and PM-2 for query number 1007, 'women in Saudi Arabia.'

The PM-2 results (bottom right) use the overall proportionality of the sentiments in the list as a guidance for choosing further documents: here, a second positive document is pulled up from lower ranks, yielding the best CPR@10 score among the 4 models for this query at a cost of slightly lower Precision-IA@10 than SCSF. With 5 negative documents, 3 mixed/neutral ones, and 2 positive documents we are very close to the desired distribution of sentiments.

3.3 Summary

In this chapter we introduce different measures at the topic level to characterize the opinionatedness of a controversial piece of text, and analyze them on the TREC Blog Track dataset. Then, we demonstrate how to diversify search results according to sentiments by considering a pre-defined bias. This allows us to emphasize either majority or minority sentiments during diversification, or to give an unbiased representation across all sentiment aspects. For this, we introduce several diversification models that use sentiments and query sentiment aspects distributions. Diversifying the output of a strong retrieval baseline, the results on the TREC Blog Track data reveal that the proportionality-based methods and the SCSF model perform best according to most measures, but an individual choice should be made based on the quality of the sentiment classifier at hand. Finally, we demonstrate the value of using biases and accurate sentiment classification for query sentiment aspects distribution estimations by means of cross-bias experiments in which equal diversification is performed instead of the desired bias.

CHAPTER 4

TEMPORAL DIVERSIFICATION

In this chapter, we focus on the temporal dimension by extracting time intervals from *within* documents. First we introduce different kinds of temporal queries in Section 4.1. Then, in Section 4.2, we describe the extraction process of time information from Wikipedia and blog documents and analyze how well they cover each other. This leads to the definition of a time measure 4.3. Finally, in Section 4.4 we present modified components of the diversification models from Chapter 3, which is followed by experimental results.

4.1 Introduction

As discussed in Section 2.4.2, there is a lot of research in temporal information retrieval. In this thesis, we focus on *topic-specific* temporal characteristics that are important for the query in question only. Users are often interested in temporal aspects of a query. An initially retrieved list of search results may not be temporally biased or it may only prefer recent search results, i.e., it may not include information from those time intervals that are important for the query but lie in the past. Typical user intents for such a situation would be 'What all important information is out there on this topic? How did the information change over time?', or 'How were the specifics for this topic, when the product or issue first emerged versus now?'. These are complex information needs that would require a user to proceed with multiple searches on the web, to then merge the relevant information together. If we knew the past trends and spiking times for a query or topic in the past, then we could choose

results from these spiking times for diversification. This prior temporal information about a query can easily be obtained by analyzing queries or click through data in a query log, or even relevance judgments in a corpus.

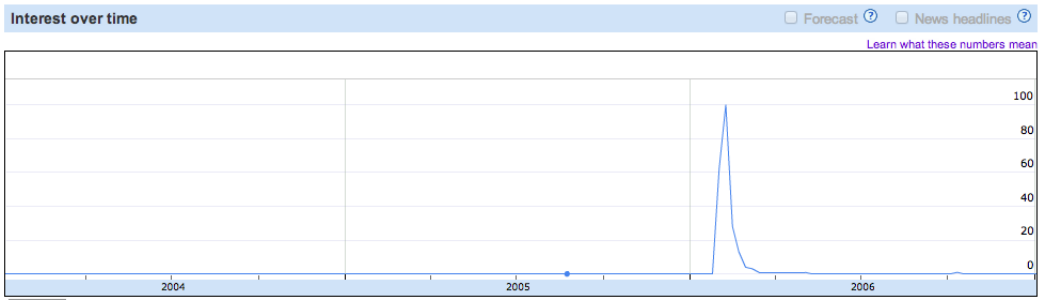

Figure 4.1. Search activity on Google during 2004-2006 for the query 'muhammad cartoon' according to Google Insights for Search.

In the following we show the search activity for a popular search engine over time for a query. Figure 4.1 shows a graph from Google Insights for Search[1] for the topic 'muhammad cartoon', which reveals high search activity around the time when the cartoon event happened. This query is an example of a temporally unambiguous one, as Jones and Diaz (2007) define them. Though we can still study diversity for such a query, most of the important documents will certainly be from the time interval starting at the end of 2005 until the end of 2006. We have another example with the query 'super bowl ads' in Figure 4.2. This is a temporally ambiguous query in that it spikes every year at a certain time. There is a lot of opportunity for diversification here – for example we could boost documents at spiking points, both in maxima as well as in minima.

Finally, we look at a truly ambiguous query that also has many spiking points and overall greater maintained popularity according to Google (Figure 4.3): 'Windows Vista'. Again, we can choose many local maxima for diversification time intervals.

[1]http://www.google.com/insights/search/

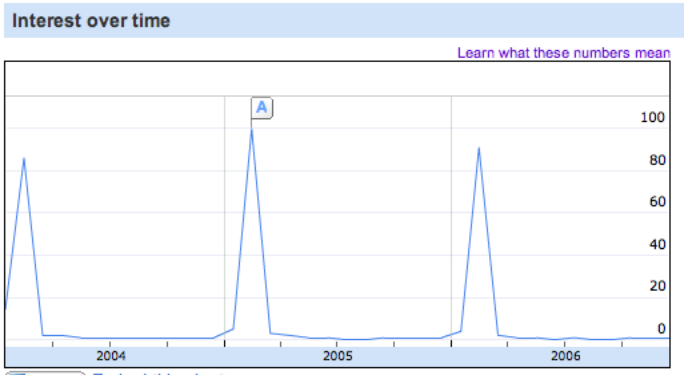

Figure 4.2. Search activity on Google during 2004-2006 for the query 'super bowl ads' according to Google Insights for Search.

Figure 4.3. Search activity on Google during 2004-2006 for the query 'windows vista' according to Google Insights for Search.

Temporally ambiguous queries are certainly the most interesting to study together with diversity, but the other classes should not be excluded. Beginning with a general query that does not contain any time references, we can diversify search results after analyzing past trends. Queries containing explicit time references like "in 2005" or "last year" require the search results to be from a predefined time frame. So this restricts our flexibility in diversification. A temporally unambiguous query may also narrow the diversification time interval, but it is still our choice to stretch and extend that interval as required. In the TREC Blog Track, we analyzed the 150 queries for temporal ambiguity by looking at Google Insights for Search results from 2004 until

2012, and only 19 of the queries were identified as temporally unambiguous. The remaining 131 queries are temporally ambiguous.

How do we identify relevant times in documents? We can either use document publication dates or we can extract times mentioned within documents with a date/time tagger. The latter has the advantage that often *time intervals* spanning longer periods than just a single day can be considered; further, pre-internet times can be included as well. Finally, a document written in the present may refer to events in the past, and hence the publication date may not accurately reflect the timing of those events. For these reasons, in this work we deal with times extracted from document content only.

4.2 Extracting Spiking Times from Wikipedia

For our diversification experiments and further analyses we choose Wikipedia as a source for extracting spiking times for queries. For this, we index an available recent full dump of Wikipedia (January 2012) with Indri. We use the 150 queries from the TREC Blog Track (Ounis et al., 2006; Macdonald et al., 2007) as in previous experiments throughout this thesis: the queries' stopped title and description texts are combined for use with the Sequential Dependence Model in Lemur/Indri (Metzler & Croft, 2005), smoothed using Dirichlet ($\mu = 10,000$). We retrieve the top 2 documents from Wikipedia for each query. Usually, for spiking date/time extraction it is sufficient to only use the corresponding Wikipedia page for a query, however this does not apply to queries for which there is no such exact page. Using more than 2 documents affects the quality of extracted times for most queries. This is why we extract times from the top 2 retrieved documents for all queries.

For date/time extraction we use the Stanford NE Tagger (Finkel et al., 2005). Date/time tags are extracted and transformed to the format "YYYY-MM-DD/ YYYY-MM-DD" using a series of regular expressions. The first date point refers

to the beginning of the time interval, whereas the second one refers to the end of the time interval similar to prior work (Berberich et al., 2010; Kanhabua & Nørvåg, 2010). This representation is most suitable for describing both time points and time ranges. To give an example, January 26, 2005 would be converted into "2005-01-26/2005-01-26", whereas the year 2004 is converted into "2004-01-01/2004-12-31". These extracted time intervals from Wikipedia are used as our diversification bins in Section 5.4, across which variety in search results shall be achieved. Further, we apply the same procedure for extracting times from retrieved blog documents. For the current experiment, we apply this date extraction procedure to the top 50 documents retrieved from the TREC Blog Track corpus with the above-mentioned 150 queries.

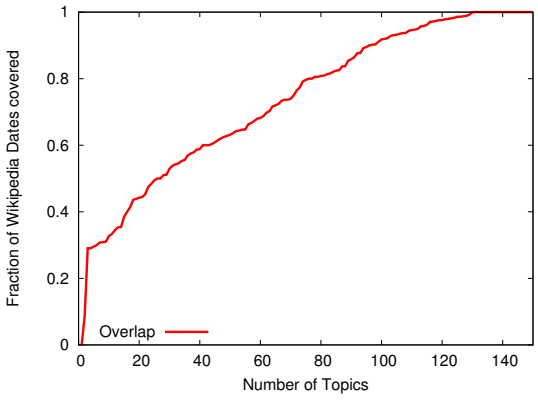

Figure 4.4. Recall: Distribution of overlap between dates extracted from top 50 Blog-retrieved documents for dates in Wikipedia.

In the following experiment we determine the degree of overlap of extracted times between the document blog corpus and Wikipedia. In particular, we want to know to what extent the times found in the documents cover those found in Wikipedia – our spiking times, representing *recall*. We analyze the data without filtering any low frequency times extracted from Wikipedia or the documents. We interpret two date intervals as 'overlapping' if they intersect for at least one day. Figure 4.4 shows the recall results sorted in increasing order of overlap with each TREC topic as a point on the x-axis. Only around 30 topics cover fewer than 50% of the Wikipedia dates,

the remaining topics achieve a good coverage. Table 4.1 shows the TREC topics at extreme ends – those that achieve poorest coverage at the top, and those achieving perfect coverage at the bottom. It becomes immediately evident that more popular controversial topics achieve better recall for Wikipedia, whereas those with poor coverage are about less discussed and less sensitive topics such as product reviews. On average the percentage of overlap in Figure 4.4 is 73.89%. If we increase the number of retrieved documents in the blog corpus from 50 to 1000 for date extraction, we are able to cover an average of 98.81% of times extracted from Wikipedia.

Table 4.1. Recall: TREC topics whose retrieved documents cover spiking times from Wikipedia poorly (top) and perfectly (bottom).

Topic	*Title*	*Overlap*
1023	Yojimbo	0.000
925	mashup camp	0.087
1035	Mayo Clinic	0.290
926	hawthorne heights	0.290
883	heineken	0.295
934	cointreau	0.300
1039	The Geek Squad	0.308
...	...	...
1007	women in Saudi Arabia	1.000
1017	Mahmoud Ahmadinejad	1.000
1019	China one child law	1.000
878	jihad	1.000
889	scientology	1.000
896	global warming	1.000
897	ariel sharon	1.000
912	nasa	1.000

At a second glance, it appears unusual to have quite a few controversial topics with perfect coverage. A quick inspection reveals that for TREC topics with excellent coverage users often quoted Wikipedia articles in blog posts and comments. Also, the blog corpus is from 2006 with TREC topics from 2006-2008, whereas the Wikipedia dump is slightly newer. Some of the newer dates in Wikipedia are covered with noisy dates in the blogs referring to the future. This is due to our flexible date matching

convention in this analysis. To deal with this, we conclude that for the experiments it is beneficial to exclude dates beyond 2009 from both the blogs and Wikipedia to have both refer to the same time ranges.

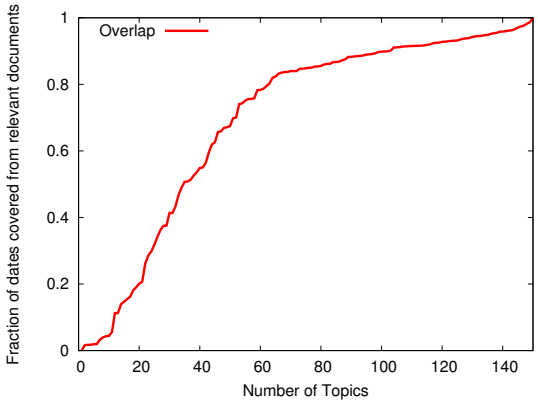

Figure 4.5. Precision: Distribution of overlap between dates extracted from Wikipedia for dates from all judged relevant blogs.

To confirm a decent match between Wikipedia and blog corpus dates, we conduct a similar analysis from a different angle, representing *precision*: how well do Wikipedia dates cover dates in judged relevant documents? In this direction the analysis is noisier since blog documents – even if judged relevant – are not as cleanly written as Wikipedia articles, and thus contain a higher number of noisy dates. The results are interesting: we have an average overlap of 69.9%. The topic-by-topic analysis is shown in Figure 4.5. The graph looks similar to Figure 4.4, but with only one perfect topic and a few more topics for which fewer than 50% of the dates are covered. A positive observation is that the line is steeper than in Figure 4.4. In Table 4.2 we again show some TREC topics with the worst covered dates (top) and those with best covered dates in relevant documents. The results look similar to Table 4.1, with the difference that the best-covered topics are not necessarily highly controversial. A quick lookup reveals that those highly controversial topics having perfect scores in Table 4.1 are also covered decently well here, in the 60%-97% range. Figure 4.6 is a combined graph, showing that precision and recall for the Wikipedia date overlap are

decently correlated. Overall, we can conclude that Wikipedia dates and blog corpus dates overlap well when matching is performed in a flexible manner.

Table 4.2. Precision: TREC topics with dates from Wikipedia covering dates in relevant documents poorly (top) and well (bottom).

Topic	*Title*	*Overlap*
1023	Yojimbo	0.000
909	Barilla	0.016
1021	Sheep and Wool Festival	0.017
939	Beggin Strips	0.018
894	board chess	0.019
950	Hitachi Data Systems	0.020
913	sag awards	0.032
...	...	...
888	audi	0.962
1044	talk show hosts	0.966
1004	Starbucks	0.970
895	Oprah	0.974
921	Christianity Today	0.976
858	super bowl ads	0.982
1008	UN Commission on Human Rights	0.987
1013	Iceland European Union	1.0

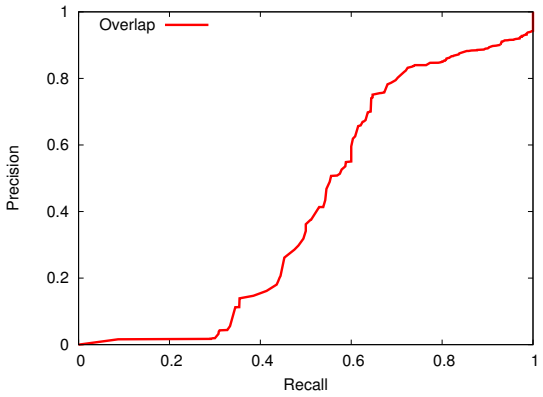

Figure 4.6. Correlation between Precision and Recall for measuring the overlap between document and Wikipedia dates.

4.3 Measures

Given is a query Q and query time aspects in the form of spiking times $\pi \in time(Q)$, such as extracted from Wikipedia. We can ask a simple question: how should retrieved documents for Q be reranked in order of relevance to these times? For this we devise a simple reranking model or measure:

$$P(D|time(Q)) = \sum_{\pi \in time(Q)} P(D|\pi) \cdot P(\pi|Q) \tag{4.1}$$

where $P(\pi|Q)$ signifies the importance of time π for Q, as given by frequencies from Wikipedia, for instance. $P(D|\pi)$ describes the likelihood of D being relevant to time π, defined as follows:

$$P(D|\pi) = \sum_{\kappa \in D} P(D|\kappa) \cdot P(\kappa|\pi) \tag{4.2}$$

where κ are times mentioned in D. $P(\kappa|\pi)$ expresses how well π and κ match. κ is a time from D and π is an important time for this query that was identified as a query time aspect earlier. We estimate $P(D|\kappa)$, the likelihood of D being relevant to time κ, as follows by applying Bayes' Rule:

$$P(D|\kappa) = \frac{P(\kappa|D) \cdot P(D)}{P(\kappa)} \stackrel{rank}{=} P(\kappa|D) \tag{4.3}$$

which is rank-equivalent by assuming that $P(D)$ is a constant and $P(\kappa)$, the prior probability of a particular time κ, is equal across all times. Hence we continue with Equation 4.2:

$$\sum_{\kappa \in D} P(\kappa|D) \cdot P(\kappa|\pi) = \sum_{\kappa \in D} \frac{c(\kappa, D)}{\sum_{\mu \in D} c(\mu, D)} \cdot \frac{|\kappa \cap \pi|}{|\kappa \cup \pi|} \tag{4.4}$$

of which the first component determines how important κ is as opposed to other times in D by means of normalized frequency counts. However, to precisely capture the

quality of the match, we also need to consider its size in number of days, which is what the second component $P(\kappa|\pi) = \frac{|\kappa \cap \pi|}{|\kappa \cup \pi|}$ determines. Berberich et al. (2010) mention three criteria that such an overlap must fulfill: specificity, coverage and maximality. We ensure specificity by normalizing extracted dates as precisely as possible. Coverage is given by discounting with $\frac{|\kappa \cap \pi|}{|\kappa \cup \pi|}$, and maximality is also ensured since $\frac{|\kappa \cap \pi|}{|\kappa \cup \pi|} = 1.0$ if $\kappa = \pi$ exactly. Clearly, this measure ranks documents matching multiple of the more important spiking times highest, and those matching less important or fewer ones lower.

4.4 Diversification

In the rest of this chapter we present *modified* parts of the models and frameworks from Chapter 3 for the time dimension. We do not change the models per se; merely the components estimating sentiments are swapped with those estimating time aspects. Then, we perform a small set of experiments solely for the time dimension to demonstrate the effectiveness of the models under different circumstances. A more extensive evaluation of the time dimension is carried out together with sentiments in Chapter 5.

4.4.1 Models

For the time dimension, diversification is performed across query time aspects $\pi \in time(Q)$ instead of query sentiment aspects $\sigma \in sent(Q)$. Each query has a variable but finite number of times t_i with weights associated with it, i.e., $time(Q) = \{t_1, \ldots, t_n\}$. Similarly, each document has a fractional t_i score so that they sum to 1.0 for a single document across all query time aspects $t_1, \ldots, t_n \in time(Q)$.

4.4.1.1 Retrieval-Interpolated Diversification

4.4.1.1.1 Time Contribution by Strength (TCS) We rename the model 'Sentiment Contribution by Strength (SCS) from Section 3.2.2.2.1 to 'Time Contribu-

tion by Strength' (TCS). This model accepts fractional non-topical query aspects for the estimation of $P(\sigma|D)$. For time, we refer to it as $P(\pi|D)$: query time aspects $\pi \in time(Q)$ are variable across queries with differing granularities for single time aspects, such as days, months, and years. Given document D and time mentions $\kappa \in D$ in form of intervals (see Section 4.4.2.2 for more detail), we estimate the relevance of D to query time aspect π as follows:

$$\begin{aligned} P(\pi|D) &= \sum_{\kappa \in D} P(\pi|\kappa) \cdot P(\kappa|D) \\ &= \sum_{\kappa \in D} \frac{|\kappa \cap \pi|}{|\kappa \cup \pi|} \cdot \frac{c(\kappa, D)}{\sum_{\mu \in D} c(\mu, D)} \end{aligned} \tag{4.5}$$

where $P(\kappa|D)$ represents the likelihood of time κ occurring in D, estimated through normalized frequency counts for those time mentions. $P(\pi|\kappa)$ expresses how well κ covers π. Given two time intervals, π and κ, $P(\pi|\kappa)$ determines whether they overlap and how significant this overlap is. A 1:1 time interval matching thus yields a score of 1.0, and non-matching times yield 0. Any other imperfect matching score lies in between these two extremes. For example, *March 10-15* is a partial match for *the month of March* and yields a score of $\frac{6}{31}$.

4.4.1.1.2 Time Contribution by Strength and Frequency (TCSF) This is analogous to 'Sentiment Contribution by Strength and Frequency (SCSF) from Section 3.2.2.2.2. Here, we swap $P(\bar{\sigma}|S)$, the likelihood of S not having sentiment σ, with

$$P(\bar{\pi}|S) = \frac{time(\bar{\pi}, S)}{|S|} \tag{4.6}$$

which is the fraction of documents in S not covering time π to at least 50%. For estimating coverage between times $\kappa \in D \in S$ and π we use $\frac{|\kappa \cap \pi|}{|\kappa \cup \pi|}$ as in Equation 4.5.

We set $P(\pi|S) = 0$ if $S = \emptyset$ to avoid zero division in the first iteration. The remaining parts of the model are straightforward sentiment vs. time swaps.

4.4.1.2 Diversity by Proportionality

The only modifications for this model are straightforward sentiment versus time aspect swaps. Fractional time scores $P(\pi|D)$ are estimated as described in Equation 4.5.

4.4.1.3 Favoring Different Biases in Search Results

For the time dimension, $P(\pi|Q)$, is estimated slightly differently depending on the bias, as explained below.

4.4.1.3.1 Equal Time Diversification (EQ) We rename the BAL bias from Section 3.2.3.1 to Equal (EQ) for time diversification for clarity. It is calculated as

$$P(\pi|Q) = \frac{1}{|time(Q)|} \tag{4.7}$$

which is a uniform distribution over all query time aspects $\pi \in time(Q)$.

4.4.1.3.2 Diversifying Towards the Query Time Distribution (SPK) The CRD bias is renamed to Spike (SPK). For sentiments, this was estimated as $P(\sigma|Q)$ – the fraction of documents having the most confident sentiment class as described in Section 3.2.3.2. For time, the 1:1 mapping of times to documents is less appropriate since we use several times mentioned within documents (see Section 4.4.2.2), many of which are often equally relevant. Therefore, given some time-tagged data for Q in the form of documents, we calculate $P(\pi|Q)$ with the normalized occurrence frequency of π in the pool of times extracted from these documents. This yields $P(\pi|Q)$, the likelihood of time π to be drawn from Q's time aspects distribution.

4.4.1.3.3 Diversifying Against the Query Time Distribution (SLB)

We rename OTL to Slab (SLB) for time diversification. We choose "slab" as a descriptive label since the tail of time distributions is usually very flat. This bias is calculated from the SPK bias via value swapping in the distribution the same way that OTL is derived from CRD (Section 3.2.3.3), resulting in a boost of the tail times. We note that with a small set of query aspects – such as for sentiments – the value of the OTL bias may be less obvious. But for dimensions such as time, where queries typically have a larger set of query time aspects, diversifying against the query time aspects distribution may be more valuable such as with the SLB bias.

4.4.2 Experimental Setup

4.4.2.1 Data

Retrieval Corpus We use the TREC Blog Track data 2006-2008 (Ounis et al., 2006) as retrieval corpus for all our experiments. For preparation, the DiffPost algorithm is applied for better retrieval as shown in prior work (Lee et al., 2008; Nam et al., 2009). Further, we perform stop word removal and Porter stemming. This is analogous to the setup of the corpus in Chapter 3.

Queries and Retrieval Model We split the 150 TREC Blog Track 2008 queries into 3 non-overlapping randomly chosen sets of size 50 each in order not to bias training or testing towards a specific year: split 1 is used for training and tuning parameters; the results in this work are reported on split 2. Split 3 is not used in this chapter, as it was reserved for sentiment classifier training. For our diversification experiments, we use a strong retrieval baseline as in Chapter 3: the queries' stopped title and description texts are combined for use with the Sequential Dependence Model in Lemur/Indri (Metzler & Croft, 2005), smoothed using Dirichlet ($\mu = 10,000$). All diversification models are applied to the top $K = 50$ retrieved documents as

determined during training. The retrieval scores are normalized to yield document likelihood scores.

4.4.2.2 Time

Extracting Times From Documents Corpora like the TREC Blog Track corpus (Ounis et al., 2006) come with metadata such as permalink and publication dates. However, this would mostly restrict times to be in the 2000s for this work. To have a more suitable and realistic setting for our task, we use document content dates. For extracting those times we use the Stanford NE Tagger (Finkel et al., 2005). By means of a series of regular expressions, tagged date portions are extracted from documents and normalized to date ranges of the form "YYYY-MM-DD/YYYY-MM-DD" as in prior work (Berberich et al., 2010; Kanhabua & Nørvåg, 2010) – indicating the beginning and end of a time mention. To give an example, January 26, 2005 is converted into "2005-01-26/2005-01-26", whereas the year 2004 is converted into "2004-01-01/2004-12-31". Ambiguous time mentions that cannot be normalized into this format are dropped. The tagger also finds documents with dates referring to the future. However, given that the TREC Blog Track corpus and queries are from 2006-2008, particularly dates beyond 2009 in the corpus are noisy. Therefore, we omit dates beyond 2009 for the experiments as a rational decision to ensure high quality.

Truth Judgments The TREC 2008 Blog Track judgments do not contain time-specific judgments. To obtain this, first for each TREC topic, times are extracted from the topic's TREC-judged relevant documents followed by normalization as described above. 47,311 date ranges are extracted in total, of which 9763 are then manually judged by an undergraduate student in psychology and a graduate student in computer science. The judgments are performed at a topic-level because relevant times for a query or topic are universal. Date ranges for a topic are judged in order of frequency until the annotator is confident that the most important relevant times

for a query have been covered. As judgment criteria we consider: a time mention is relevant to the query if it refers to a relevant event. The time at which an opinion was uttered or a document was published is considered non-relevant *unless a relevant event is associated with that time.* The obtained judgments are then used for two purposes: (1) **as document-level time judgments**: converted back to the document level they serve as ordinary binary "subtopic" or query aspect judgments. The 9763 topic-level judgments translate into 178,678 document-level time judgments for all topics; (2) **as Truth Query Time Aspects Distribution Estimation**: for each query we use the relevant judged times and their weights as aggregated occurrences in all relevant documents for the query to estimate the Truth Query Time Aspects Distribution.

Query Time Aspects Estimation For the experiments, the query time aspects (variable per query) and their distribution are estimated from Wikipedia. For this, we index a recent full dump of Wikipedia (January 2012) with Indri, from which links and references are removed to retain only textual content. This way, times from references do not influence the calculations. Using the queries for the experiments, we retrieve the top 2 documents and process the dates from the Wikipedia pages as described above: with only 1 document retrieved, topics that do not have an exact matching Wikipedia page are at a disadvantage, and using more than 2 documents yields noisy time aspects as determined during training. The extracted and normalized Wikipedia dates and their normalized frequencies are then used as time diversification bins and query time aspects distributions during the experiments. Only dates until 2009 are extracted to match the time range of the Blog corpus, as explained earlier. Finally, we use the train split to determine a (diversification) algorithm-specific threshold to use the top x% of obtained dates per query as time bins. Good thresholds are 90% or 100%, which shows that the inclusion of less frequent extracted dates proves useful.

4.4.2.3 Evaluation Measures

We use the same style of evaluation and measures as described in Section 3.2.4.2.

4.4.3 Results

In this section we discuss the results of the retrieval baseline SDM and the diversification models proposed in Section 4.4.1, TCS, TCSF, PM-2, and PM-2M, with the three target biases, Spike (SPK), Equal (EQ), and Slab (SLB). The interpolation parameter $\lambda \in \{0.0, \ldots, 1.0\}$ is tuned in 0.1 steps separately for each model and bias on our training split. The results are presented with fixed parameters K and λ on the test split 2, and the evaluation is performed with our time judgments at rank 20 (see Section 4.4.2.2). For α-NDCG, ERR-IA and NRBP we set $\alpha = \beta = 0.5$. Statistical significance tests are reported using the paired two-sided t-test with p-value < 0.05; smaller p-values are explicitly stated with the results.

4.4.3.1 Straight-Bias Experiments

With the Straight-bias experiments our aim is to judge diversification performance by using the same bias during experiments and evaluation, similar to Section 3.2.4.3.1. For these experiments we use the Wikipedia query time aspects and weights, and evaluate with our time judgments. The results are shown in Table 4.3. Note that we are using a different set of judgments than in Chapter 3, and therefore the results are not directly comparable.

There are some interesting trends: with the SPK bias we note the best performance across all three target biases, with significantly better results for some measures over SDM and in some cases over TCS for TCSF and the proportionality models. For EQ and SLB only a few α-NDCG@20, ERR-IA@20, and CPR@20 results are significantly better over the SDM baseline. In Section 4.2 when we analyzed Wikipedia time aspects, we found a good overlap between those and the times from TREC-judged relevant documents. However, the results in Table 4.3 indicate that the weights for

the Wikipedia estimated time aspects do not agree well with those from the Truth Query Time Aspects Distribution. Both distributions are very flat as many times occur similarly frequently in relevant documents, and although the most frequently occurring times take up larger portions of the distribution, these are still small overall. As a consequence of this, the situation here is similar to having a poor sentiment classifier as in Chapter 3. As we will see in Section 4.4.3.2, however, we can significantly improve the results for the Slab bias by collapsing times and their weights, and later in Section 4.4.3.3 we show the maximum achievable performance with our time judgments. As we will see in Chapter 5 in the Straight-Bias results when the sentiment and time results are combined, the sentiment dimension is able to compensate for the weaker time results we observe here. Finally, note in Table 4.3 that although TCS does not yield significantly better results than the other methods with most measures, in some situations it performs better: with the EQ and SLB biases, TCS performs best for Precision-IA@20, s-recall@20, and for CPR@20. This hints at TCS being more compatible with dimensions having many aspects, particularly those with flatter weight distributions as observed here.

Measure	Bias	Precision-IA@20	s-recall@20	α-NDCG@20	ERR-IA@20	CPR@20	NRBP
SDM	SPK	0.2893	0.4631	0.5598	0.4682	0.8793	0.4075
TCS	SPK	0.2904	0.4713	0.5686	0.4630	0.8841	0.3921
TCSF	SPK	**0.3120***	0.4510	0.5638	0.4635	0.8900	0.3964
PM-2	SPK	**0.3134***	0.4587	**0.5893***	**0.4816**	**0.8925**	0.4088
PM-2M	SPK	**0.3142***	0.4467	**0.5809**	**0.4773**	0.8862	**0.4067**
SDM	EQ	0.1734	0.4631	0.3950	0.3111	0.8640	0.2578
TCS	EQ	0.1846	0.4840	**0.4321**	0.3224	0.8828	0.2502
TCSF	EQ	0.1827	0.4668	0.4184	0.3155	0.8825	0.2481
PM-2	EQ	0.1802	0.4686	**0.4359**	**0.3302**	0.8695	0.2588
PM-2M	EQ	0.1820	0.4546	**0.4301**	**0.3285**	0.8703	0.2584
SDM	SLB	0.1063	0.4631	0.2687	0.1976	0.7315	0.1522
TCS	SLB	0.1186	0.4787	**0.3207**	0.2174	0.7572	0.1524
TCSF	SLB	0.1153	0.4681	**0.3098**	0.2164	**0.7568**	0.1582
PM-2	SLB	0.1176	0.4645	**0.3283**	**0.2293**	**0.7547**	0.1623
PM-2M	SLB	0.1150	0.4521	**0.3063**	**0.2141**	0.7348	0.1528

Table 4.3. Straight-Bias results for all measures. Bold entries are significantly better than the SDM baseline (p-value < 0.02), whereas bold and starred entries yield a significant gain over TCS (p-value < 0.04).

4.4.3.2 Collapsing Dates for Query Time Aspects

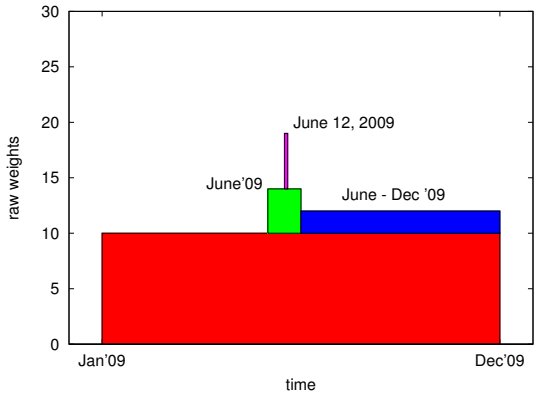

Figure 4.7. Some collapsed time intervals for the topic '2009 Iranian presidential election'.

For the query time aspects and truth query time aspects estimation we did not modify the date units or weights obtained from the judgments or extraction process. However, since many times of different granularities can be associated with a single query, and some of these may overlap, we also tried a variation of the experiments by collapsing overlapping dates and their weights. To give a non-TREC Blog Track example, for the topic '2009 Iranian presidential election' the year 2009 is relevant, but so is June 2009 as a separate time when the elections and related events took place, as well as the actual election day in that month. We can consider the importance of these times independently of each other, or overlapping times can boost each others' weights by collapsing them. We visualize collapsing weights for this example in Figure 4.7: let year 2009 have an initial unnormalized weight of 10 units, visualized in red and let June 2009 (green) have weight 4. Further, the day of the elections, June 12, 2009, is assigned a weight of 5, shown in pink. And finally, the second half of 2009 in dark blue has a raw weight of 2. By collapsing these query time intervals we can stack the weights of overlapping intervals: June 12, 2009 gets a total weight of $10 + 4 + 5 = 19$, June 2009 gets 14, and the second half of 2009 gets 12. This way, the distribution of query time weights is slightly altered to emphasize smaller time units. The basic

idea is that weights of larger time intervals are passed on to smaller overlapping time intervals. If the intersecting time interval is not already included in our set of times, it is added and inherits the weights of the intersecting 'parent' times. The weights are renormalized to sum to 1.0 after all times are processed.

SLB	**Precision-IA@20**	**Improvement**
SDM baseline	0.1447	+36.1%
TCS	0.1711	+44.3%
TCSF	0.1637	+42.0%
PM-2	0.1567	+33.2%
PM-2M	0.1619	+40.8%
SLB	**α-NDCG@20**	**Improvement**
SDM baseline	0.3334	+24.1%
TCS	0.3781	+17.9%
TCSF	0.3703	+19.5%
PM-2	0.3821	+16.4%
PM-2M	0.3718	+21.4%
SLB	**CPR@20**	**Improvement**
SDM baseline	0.7968	+8.9%
TCS	0.8188	+8.1%
TCSF	0.8108	+7.1%
PM-2	0.8124	+7.6%
PM-2M	0.8018	+9.1%

Table 4.4. Results with collapsed dates for SLB with relative improvements with respect to not collapsing dates. *All* entries are significantly better than their counterpart non-collapsed results (p-value < 0.006).

In terms of the experiments, collapsing dates only helps the results for the Slab bias, shown in Table 4.4. Over all measures and approaches, we note significant improvements with p-value < 0.006 as opposed to not collapsing dates. The gains are huge for Precision-IA@20, great but smaller for α-NDCG@20, and smallest for CPR@20. As a contrast to these results, in Table 4.5 we present the results for the Spike bias: here we have significant losses over all but one result when dates are collapsed. For the Equal bias we note small differences between the two approaches. Why does collapsing dates only help the Slab bias? This bias focuses on the tail

SPK	Precision-IA@20	Loss
SDM baseline	0.2281	-21.2%
TCS	0.2477	-14.7%
TCSF	0.2318	-25.7%
PM-2	0.2379	-24.1%
PM-2M	0.2411	-23.3%
SPK	**α-NDCG@20**	**Loss**
SDM baseline	0.4873	-13.0%
TCS	0.5082	-10.6%
TCSF	0.5144	-8.8%
PM-2	0.5285	-10.3%
PM-2M	0.5263	-9.4%
SPK	**CPR@20**	**Loss**
SDM baseline	0.8646	-1.7%
TCS	**0.8732**	-1.2%
TCSF	0.8718	-2.0%
PM-2	0.8708	-2.4%
PM-2M	0.8695	-1.9%

Table 4.5. Results with collapsed dates for SPK with relative losses with respect to not collapsing dates. *All* entries *except* for the bold ones are significantly worse than their counterpart non-collapsed results (p-value < 0.04).

distribution of times for the query. However, for the time dimension, the tail is often very flat and sparse. By collapsing dates, important small-interval time ranges are more strongly emphasized, which seems to help in the experiments. This does not make a noticeable difference for the Spike bias though. A reasonable explanation is that for most queries the front of the distribution has times with larger intervals, so those do not heavily profit from collapsing times and weights.

4.4.3.3 Perfect Time Aspects

In this style of experiment we perform straight-bias experiments for the time dimension with perfect ("oracle") time aspects and weights during experiments and evaluation by using our time judgments. The results are shown in Table 4.6. As opposed to Table 4.3, we unsurprisingly see significantly better results for almost all methods and almost all the measures over the SDM baseline, and in some cases over

TCS. Unlike our observations for SCS in Chapter 3 and DCS later in Chapter 5, here TCS sometimes achieves the best results: particularly for the Spike and Equal biases, such as for the measures s-recall@20, α-NDCG@20, and ERR-IA@20. We hypothesize that TCS may be a good approach to use for dimensions with many query aspects such as with the time dimension. However, it performs rather poorly for dimensions with a small number of aspects and a very skewed distribution, such as with sentiments. This is an interesting question to explore for future work when reliable data for other dimensions becomes available.

Measure	Bias	Precision-IA@20	s-recall@20	α-NDCG@20	ERR-IA@20	CPR@20	NRBP
SDM	SPK	0.2893	0.4631	0.5598	0.4682	0.8793	0.4075
TCS	SPK	**0.3772**	**0.5381**	**0.6502**	**0.5383**	**0.9691**	**0.4649**
TCSF	SPK	**0.4071***	0.4502	**0.6399***	**0.5367**	**0.9725**	**0.4699**
PM-2	SPK	**0.3959**	**0.5007**	**0.6321***	**0.5202**	**0.9736**	**0.4461**
PM-2M	SPK	**0.3857**	0.4878	**0.6030**	**0.4906**	**0.9577**	0.4198
SDM	EQ	0.1734	0.4631	0.3950	0.3111	0.8640	0.2578
TCS	EQ	**0.2481**	**0.5539**	**0.5075**	**0.3832**	**0.9611**	**0.2985**
TCSF	EQ	**0.2369**	0.4962	**0.4937**	**0.3797**	**0.9572**	**0.3010**
PM-2	EQ	**0.2341**	**0.5433**	**0.4899**	**0.3659**	**0.9348**	**0.2812**
PM-2M	EQ	**0.2243**	**0.5067**	**0.4560**	**0.3429**	**0.9198**	0.2642
SDM	SLB	0.1063	0.4631	0.2687	0.1976	0.7315	0.1522
TCS	SLB	**0.1765**	**0.5443**	**0.4526**	**0.3378**	**0.8612**	**0.2596**
TCSF	SLB	**0.1769**	**0.5191**	**0.4496**	**0.3374**	**0.8646**	**0.2611**
PM-2	SLB	**0.1805**	**0.5224**	**0.4829***	**0.3817***	**0.8687**	**0.3010***
PM-2M	SLB	**0.1558**	**0.4979**	**0.3377**	**0.2318**	**0.7774**	**0.1628**

Table 4.6. Results with perfect time labels for all measures. Bold entries are significantly better than the SDM baseline (p-value < 0.05), whereas bold and starred entries yield a significant gain over TCS (p-value < 0.02).

4.5 Summary

In this chapter we focus on the time dimension: dates are extracted from within documents and normalized for further use. Since Wikipedia contains many time mentions for a large variety of topics, some analyses are undertaken to measure the overlap between Wikipedia time mentions and those within relevant documents in the TREC Blog Track. We conclude that Wikipedia is a suitable source for obtaining time diversification bins. Then, a time measure is presented according to which documents can be reranked in order of relevance to this dimension. Afterwards, we adapt the

diversification models presented in Chapter 3 to the time dimension. Experiments are conducted with noisy versus perfect labels, which reveals that although there is potential for large gains using our diversification models, when the query aspect weights are noisy, we note fewer gains across the models with the three extreme target biases. Additionally, we also conduct experiments with collapsed dates and weights and note that this particularly helps the Slab bias.

CHAPTER 5

INTERESTINGNESS

This chapter describes the work for "interestingness", which is a single label for several dimensions having non-topical aspects, such as opinionatedness and time. We first motivate the importance of studying these dimensions together for controversial topics and give insights using the measures provocativeness, balance, and topic sentiment, introduced in Chapter 3. In Section 5.2 we then perform an analysis on two query logs for evidence about whether users are looking for subjective and temporal search results. This is followed by the definition of an interestingness measure involving sentiments and time in Section 5.3. Then, in Section 5.4 we focus on the diversification process for interestingness with a general bias framework integrating several dimensions with non-topical aspects and different kinds of biases. Part of this chapter is in submission for publication (Aktolga & Allan, 2014).

5.1 Introduction

In Section 1.2.3, we had introduced two criteria we will consider for interestingness in this thesis – opinionatedness and time. We argue that these two dimensions are worthwhile to study together. A possible issue is that an initially retrieved list of opinionated results for a controversial query may not reflect opinions from different time intervals – particularly important ones for the topic lying in the past. Typical user intents for such a situation would be 'How do people think about this topic now versus in the past?', or 'What was the general opinion on this person or product when it first emerged, or when important events happened versus now?'. These are complex

information needs that would require a user to proceed with multiple searches on the web, to then merge the relevant information together. If we knew the past trends and spiking times for a query or topic in the past, then we could choose opinionated results from these spiking times for diversification.

In the following, we study how the measures provocativeness (PROV), balance (BAL), and topic sentiment (TS) (introduced in Chapter 3) change over time for a TREC topic. We do this for the same queries/topics as in Section 4.1. Figure 5.1 shows how the measures changed over time for the query 'muhammad cartoon', as determined by means of relevance judgments in the TREC Blog Track corpus. We can clearly see the spike in all measures at the end of 2005 when the event happened. Surprisingly, later in 2006, while the sentiments remain highly provocative, they tend slightly to the positive side as revealed by the balance and topic sentiment measures. This may mean that the effect of the event wore off as time passed.

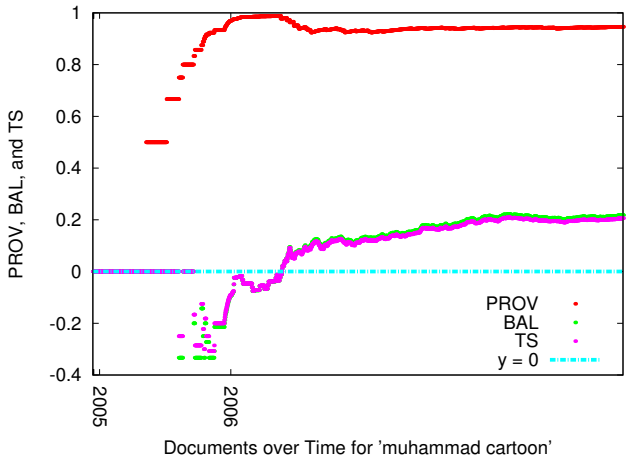

Figure 5.1. PROV, BAL, and TS for TREC topic 869 ('muhammad cartoon') arranged over time on the x-axis.

We have another example with the query 'super bowl ads' in Figure 5.2. This temporally ambiguous query spikes every year at a certain time but opinion-wise it is also very interesting, since it has many mixed opinion relevance judgments in the TREC Blog Track. In Figure 5.2 we can observe varying provocativeness around the

time the event occurred at the end of 2005, fluctuating further in 2006. While the topic sentiment seems to be balanced for most of 2005, in 2006 it tends slightly to the negative side.

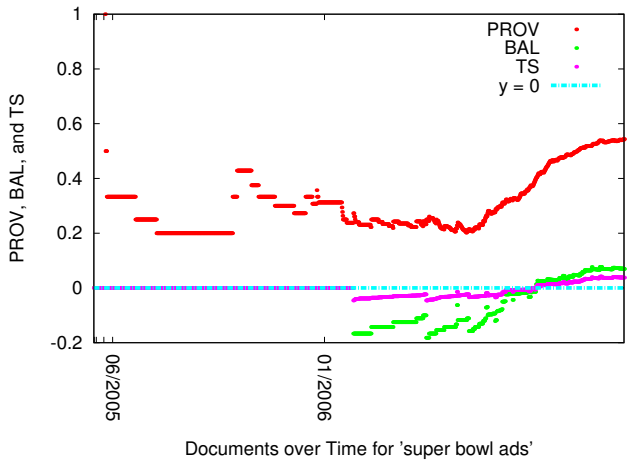

Figure 5.2. PROV, BAL, and TS for TREC topic 858 ('super bowl ads') arranged over time on the x-axis.

Finally, we look at the truly ambiguous query 'Windows Vista'. Only the time range until the end of 2006 is shown to align it with the blog track in Figure 5.3. As we can see in these examples, interesting changes occur in the measures over time.

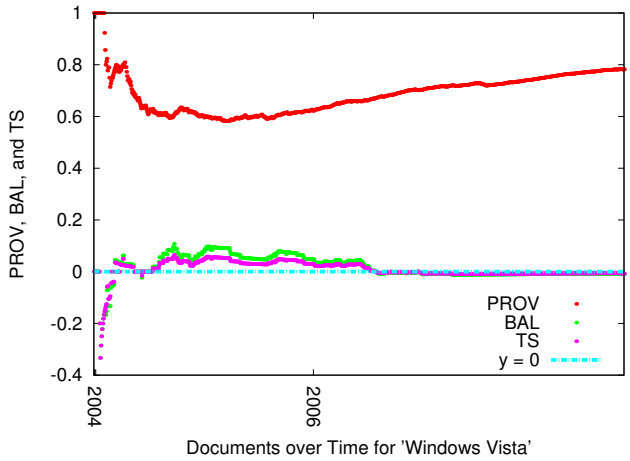

Figure 5.3. PROV, BAL, and TS for TREC topic 1005 ('Windows Vista') arranged over time on the x-axis.

5.2 Query Log Study

In this study, we analyze the MSN and AOL query logs for evidence about whether users are looking for subjective and temporal search results. Prior work (Nunes et al., 2008; Metzler et al., 2009) has extensively investigated query logs for temporal information. Metzler et al. (2009) focus on implicitly year-qualified queries. Such queries typically contain a year mention like "sigir 2008". Their analysis reveals that 7% of the queries in the query log belong to this category. In our own observation we find that the most frequently seen temporal cues in queries are years rather than completely specified dates or times. Metzler et al. (2009) present an algorithm for automatically mining such queries by observing simple patterns. Nunes et al. (2008) did a detailed study on temporal expressions in two AOL query logs, one of which we also work with in this section. Using a date/time tagger, they found a total of 1.5% or 532,989 queries including temporal expressions in the AOL query log. Removing duplicate queries, this percentage slightly increases to 1.9%. They further conclude that about 42.5% of the queries indicate a date from 2006, which is current for the query log, and 49.9% queries refer to dates prior to 2006, and the remaining 4.2% are for future dates. They further present a breakdown of query categories found.

We observe similar trends in both the AOL and MSN query logs. In general, since queries tend to be more fact-based, it is hard to explicitly find queries pointing to subjective content, but there are quite a few queries suggesting that users are looking for that: some examples are "human rights 1997", "jimmy carter human rights", and "1998 calcasieu [sic] homicides". We also found queries explicitly asking for temporal results: "happenings of 2003", "timeline of 2003", and "What's happened in 2006". There are also more verbose queries that may or may not contain specific date mentions like "what was japanese military thinking attacking the united states on December 7 1941", or "what was japan thinking when attacking paerl [sic] harbor", and "George bush broke a promise in 1991 to his voters what did he brake [sic] and

Table 5.1. Examples of reformulated controversial queries within a user's session across all sessions in both the AOL and MSN query logs. Time mentions are highlighted in bold font.

Query 1	*Query 2*
divorce christian century	divorce christian century **2006**
divorce statistics	divorce statistics **2005**
abortion statistics	abortion statistics **2005**
abortion statistics clinton vs bush	abortions **1994-2004**
what were the laws of abortion **five years ago**	what did one have to do to get an abortion in **2004**
abortion and psychology	abortion and psychology **2006**
teen abortion rate in arizona	teen abortion rate in arizona in **2000**
census bureau, illegal immigration	census bureau, illegal immigration, **2004**
ralphs supports illegal immigration protests	ralphs supports illegal immigration protests on **May 1,2006**
illegal immigration boycott	illegal immigration boycott of **May 1st**
death penalty for sexual predators in WA	death penalty for sexual predators in WA **2006**
death penalty	death penalty **2006**
Inmates aquitted from the death penalty in the **1980**'s	Inmates aquitted from the death penalty **70's and 80's**
innocent people die from the death penalty	innocent people die from the death penalty **2000**
death penalty history louisiana	death penalty history louisiana **1940**
smoking statistics since advertisments for quitting	smoking statistics **2006**
ontario no smoking laws	ontario no smoking in the workplace laws **2006**
global warming	global warming **today**
global warming	global warming in **2050**
pros of gay marriages	pros and cons of gay marriages in **2006**
gay marriage mass	gay marriage mass **2006**
gay marriages in **2005**	gay marriages in **2006**
genocide in africa	genocide in africa **2006**
why is there still genocide	why is there still genocide in the world **today**
rwanda genocide	rwanda genocide **1997**
pros and cons on euthanasia in the netherlands	pros and cons on euthanasia in the netherlands **2000**
euthanasia	euthanasia in belgium **2000**
marijuana tax act	marijuana tax act of **1937**
marijuana laws in the **1960's**	marijuana slang in the **1960s**
marijuana festivals canada	marijuana festivals canada **2006**
indian pakistan kashmir	indian pakistan kashmir **1947**
cloning and economics	cloning prohibition act of **1997**
jihadists in south africa	jihadists in south africa **today**

why". It is interesting to observe that even if these queries do not contain specific dates, they name events referring to a specific time.

A quick manual search for common well-known controversial queries yields the following instances:

- **death penalty**: gender bias in the death penalty, history on the death penalty, utah death penalty, Find me some information on the death penalty, The case against the death penalty, pros of death penalty, conservative death penalty, pro death penalty arguments, death penalty cases in israel
- **abortion**: abortion rights, abortion in the 1900, abortion view, the pros and cons of abortion

- **euthanasia**: euthanasia and its disadvantages, euthanasia and its advantages, nazi euthanasia, baby euthanasia, What are some advantages to euthanasia?
- **global warming**: timeline until global warming, causes of global warming, global warming controversy and pro and con

These are some interesting examples. For a more formal analysis, we observed the AOL and MSN query logs for 35 well-known controversial queries such as "global warming", "jihad", "osama bin laden", "genocide", "gay marriage" etc. and determined how often these queries occur as part of the query text in the logs: in the AOL query log we found 42,320 queries from a total of 36,389,567 non-unique queries in the log. So this is 0.116% of all queries. In the MSN query log we found 12,629 occurrences in a total of 14,921,285 non-unique queries in the log. That is a mere 0.085%. These numbers constitute a small proportion of the whole query log, but yet users do issue them.

As a more direct analysis of how much users are interested in the time aspect of subjective queries we observe cases in a user's session where a query is reformulated to include a time reference. Queries that already contain a time reference but are then reformulated with a different one are also included. Some examples are shown in Table 5.1, arranged by topic. It is evident that there is a wide coverage of controversial topics. Further, the queries are often very verbose and thus rare.

Since controversial queries are very sparse in the two query logs, and such queries with time references are even rarer, the question of whether users ask for multiple times for a topic is tough to answer on the basis of the query logs at hand. But still, we located some relevant examples: "shelby co il divorce records 1983-1987", "abortions 1994-2004", "genocides of the 20th century", and "chart statistics on death penalty in 2005-2006". Note that even if a user issues a query mentioning only a single year, results containing other time mentions spanning relevant days or months in the specified year are also often considered relevant. For search result ranking then the

question arises of how to rank such documents: typically, documents containing a maximum number of relevant time mentions overlapping well with the desired time should be boosted in the ranking. So diversification across times can be very useful in this situation, even if the user only indicate one time mention in her query.

Overall, we do not have any information about whether these searches were successful or satisfying. But our aim in this work is to improve search result ranking for such queries so that users can be more satisfied with what they get. We achieve this by presenting results from varying times and different sentiments in this chapter.

5.3 Measures

Given is a query Q and query time aspects in the form of spiking times $\pi \in time(Q)$. In Section 4.3, we had devised the following time measure for reranking retrieved documents in order of relevance to these times:

$$\begin{aligned} P(D|time(Q)) &= \sum_{\pi \in time(Q)} P(D|\pi) \cdot P(\pi|Q) \\ &= \sum_{\pi \in time(Q)} \sum_{\kappa \in D} \frac{c(\kappa, D)}{\sum_{\mu \in D} c(\mu, D)} \cdot \frac{|\kappa \cap \pi|}{|\kappa \cup \pi|} \cdot P(\pi|Q) \end{aligned} \tag{5.1}$$

$P(D|sent(Q))$ – the equivalent for the sentiment dimension – can be defined analogously:

$$\begin{aligned} P(D|sent(Q)) &= \sum_{\sigma \in sent(Q)} P(D|\sigma) \cdot P(\sigma|Q) \\ &\overset{rank}{=} \sum_{\sigma \in sent(Q)} P(\sigma|D) \cdot P(\sigma|Q) \end{aligned} \tag{5.2}$$

in which documents are ranked in order of relevance to sentiment by considering with $P(\sigma|D)$ how well they address different query sentiment aspects, and with $P(\sigma|Q)$, how important those sentiment aspects are for the query. $P(\sigma|D)$ can be estimated through a sentiment classifier as in Chapter 3. With this measure, controversial documents addressing the most valuable sentiments for a query will be ranked highest.

We can see that the definitions for the two dimensions have similarities. A general measure can easily be inferred for any dimension $M \in \mathcal{M}$ with countable non-topical query aspects $\omega \in asp(Q)$ as follows:

$$P(D|asp(Q)) = \sum_{\omega \in asp(Q)} P(D|\omega) \cdot P(\omega|Q) \tag{5.3}$$

where $P(D|\omega)$ indicates how well query aspect ω is fulfilled with respect to document D, and $P(\omega|Q)$ signifies the importance of aspect ω to query Q, which is typically derived from the query aspect distribution weights. As a next step we can devise a general generative definition of interestingness, according to which retrieved documents for Q can be reranked so that the most interesting documents are boosted:

$$\begin{aligned} P_{interesting}(D|Q) &= \sum_{M \in \mathcal{M}} P(D|M) \cdot P(M|Q) \\ &= \sum_{M \in \mathcal{M}} P(D|asp(Q)) \cdot P(M|Q) \end{aligned} \tag{5.4}$$

where we estimate the interestingness of D given the query Q and the interestingness dimensions $M \in \mathcal{M}$ by observing how well D fulfills the query aspects of each dimension M: this corresponds to $P(D|asp(Q))$. Further, we have a distribution over the dimensions $M \in \mathcal{M}$ for Q, so $P(M|Q)$ indicates how importance each M is for Q. Just as query aspects $\omega \in asp(Q)$ define a distribution over the aspects for Q for a single dimension M, $P(M|Q)$ defines a distribution over the dimensions $M \in \mathcal{M}$ for Q.

Our interestingness dimensions in this thesis are $M \in \mathcal{M} = \{\text{sentiment, time}\}$. So given these, Equation 5.4 can be further expanded as follows:

$$= P(D|sent(Q)) \cdot P(sentiment|Q) + P(D|time(Q)) \cdot P(time|Q) \tag{5.5}$$

Here, $P(sentiment|Q)$ and $P(time|Q)$ express the importance of each query dimension for Q. We set this to a query-dependent constant that controls the weighting between the two components. One way of defining this is

$$P(sentiment|Q) = PROV(T(Q)) = \frac{|rel(T) \setminus O|}{|rel(T)|} \tag{5.6}$$

which is the provocativeness measure for Q's topic T defined in Chapter 3. In other words, the importance for the query sentiment dimension is determined by how provocative the associated topic is. $P(time|Q)$ is then defined as $1 - P(sentiment|Q)$ accordingly. Alternatively, this constant can be tuned through a parameter sweep on a held-out dataset.

By combining sentiment and time estimation for a query in this way, $P_{interesting}(D|Q)$ produces a ranking of documents fulfilling both dimensions to the desired degree. More "interesting" documents are thus ranked higher than less interesting ones. Note that no diversity or biases are introduced in the picture yet. We move on to diversification with sentiments and time in the next section.

5.4 Diversification

5.4.1 Introduction

As described in Chapter 2, topical diversification with equal preference for all aspects has been well researched in the information retrieval community. Initial approaches to opinion diversity with an equal bias for sentiment aspects and temporal diversification were presented. In Chapter 3, we show how to do sentiment diversification with various biases. However, none of this prior work has introduced a

general framework for simultaneously diversifying across multiple dimensions with different biases. Dimensions can be of topical or non-topical nature. The latter refers to vertices under which a given topic can be studied for non-topical properties such as sentiments, time, and geography to name a few. In this chapter, we present two frameworks for diversifying with non-topical aspects and their different biases by focusing on the sentiment and time dimensions. These are essential because particularly when researching sentiments on a topic it helps to see how they evolved over time as events happened. The ability to switch between perspectives and different time emphases allows the user to understand past opinions and associated sentiments better, at what time they changed and what caused them to do so.

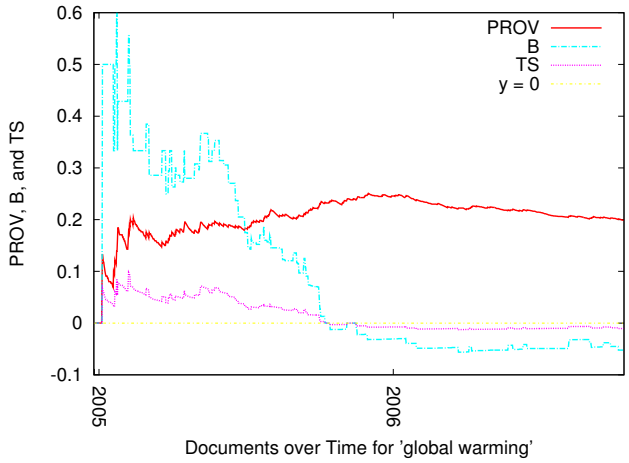

Figure 5.4. Provocativeness (PROV), Balance (B), and average sentiment (TS) values over time for 'global warming' (number 896, TREC Blog Track).

To supplement the examples presented in Section 5.1 and motivate diversification across several dimensions, consider the query 'global warming.' In a typical use case, a user engages in a comprehensive literature review with the aim of understanding the positions on this topic. This involves browsing through opinionated documents and mentally categorizing the results, for example by sentiments – positive, negative, neutral, and mixed (Kacimi & Gamper, 2012). However, opinions on a topic are often closely tied to situations and events that happened at a certain time. Con-

sider for example Figure 5.4 in which we show how sentiments change over time for 'global warming'. As in Section 5.1, we use the provocativeness, balance, and average topic sentiment measures on the TREC Blog Track data (Ounis et al., 2006) with the blog publication times. We can observe in Figure 5.4 that rather positive (i.e., unconcerned) opinions were expressed about global warming until the first half of 2005, however this trend kept declining over the second half of the year. Towards the end of 2005 the overall balance of sentiments shifted to negative with a slight upward slope in provocativeness, indicating that more opinionated blog documents were posted. The negative discussion continued throughout 2006. What happened in the middle of 2005? A quick search on the web reveals that 2005 was one of the warmest years so people got very concerned and aware about the negative issues of global warming when the hot summer came.

How do search engines handle controversial queries? Often, to be unbiased they return neutral to positive results (Demartini & Siersdorfer, 2010). Time-wise, given prior research for ordinary queries there is a strong emphasis on recent search results (Dai et al., 2011; Dong et al., 2010; Elsas & Dumais, 2010; Jatowt et al., 2005). This is an appropriate response for most user search intents. However, during tasks such as a literature review requiring more widespread information about a topic, the following would be valuable:

1. the ability to *switch the result perspective* to better grasp the polarity of opinions (Aktolga & Allan, 2013). Typical perspectives would be (a) a balanced and unbiased viewpoint; (b) a representation that emphasizes majority opinions; and (c) one that stresses minority opinions;

2. the ability to *make choices about times*: (a) results with an equal preference for all times relevant to the query; (b) results emphasizing crucial times and events for the query; (c) results emphasizing less important times for the query.

By offering such options to the user time can be saved for manually analyzing large amounts of data to understand majority/minority opinions, finding important events associated with the query, and establishing a connection between the two dimensions. Instead, the system bears the burden by analyzing a reliable source of data for the pre-existing sentiment bias for a controversial query and by extracting 'spiking' or crucial times at which important events happened. With this additional knowledge, the 'Query Sentiment Aspects Distribution' and 'Query Time Aspects Distribution', we can then diversify search results considering the user's desired biases.

In this chapter, we introduce a general bias framework for *multiple* dimensions with non-topical aspects that is integrated in two diversification frameworks. The experiments are performed on the TREC Blog Track with three extreme target biases for each dimension that are straightforward to evaluate: three sentiment biases – Balance, Crowd, and Outlier, and three time-specific biases Spike, Slab, and Equal – are used together with different diversification algorithms. We interpret the findings by observing the results from different angles.

5.4.2 Diversification Framework for Non-Topical Aspects

There are a number of ways that diversification can be modeled. The most common approaches are explicit aspect diversification (R. L. Santos et al., 2010a), which we consider in Section 5.4.2.3 or through proportionality (Dang & Croft, 2012), which we adapt in Section 5.4.2.4.

Given is a query Q and a dimension $M \in \mathcal{M}$ with non-topical countable query aspects $\omega \in asp(Q)$, across which search results will be diversified. We will use the distribution of query aspects $\{\omega_1, \ldots, \omega_n\} \in M$ for Q in our models to diversify search results across a single or multiple dimensions. Each dimension has its own variable but finite number of query aspects with different weights. This information is the

'trend' or bias that is used in the diversification frameworks below for introducing variety in a ranked list of search results.

In this work we consider the dimensions *sentiment* and *time*, so diversification is performed across query sentiment aspects $\sigma \in sent(Q)$ and query time aspects $\pi \in time(Q)$. More specifically, $sent(Q) = \{positive, negative, neutral\}$, which is constant across all queries. Each document has a fractional positivity, negativity and neutrality score summing to 1.0 for a single document (Aktolga & Allan, 2013). The time dimension is more complex in that each query has a variable but finite amount of times t_i associated with it, i.e., $time(Q) = \{t_1, \ldots, t_n\}$. Similarly, each document has a fractional t_i score so that they sum to 1.0 for a single document across all query time aspects $t_1, \ldots, t_n \in time(Q)$.

5.4.2.1 Non-Topical Biases

In the diversification frameworks below we utilize $P(\omega|Q)$, the importance of a non-topical aspect ω to query Q. Different biases can be enforced during diversification and in the evaluation by interpreting this component in different ways. We present a general bias framework applicable to any non-topical dimension and briefly detail how to compute biases for the sentiment and time dimensions. We provide two variations of the framework below. Figure 5.5 shows what this means for the sentiment dimension: given the Crowd bias for a query, we present two ways of estimating the Outlier bias from this.

5.4.2.1.1 Reverting the Distribution Given is a query Q and dimensions $M \in \mathcal{M}$. Our task is to estimate the distribution of non-topical query aspects $\omega_i \in asp(Q)$ for each dimension M. For this, let some data be given for Q that is tagged with respect to each dimension M. In this work, typically the data for Q is in the form of documents, but it could theoretically also come from a query log or query-to-concepts graph etc. So this data will be used as source for estimating the distribution

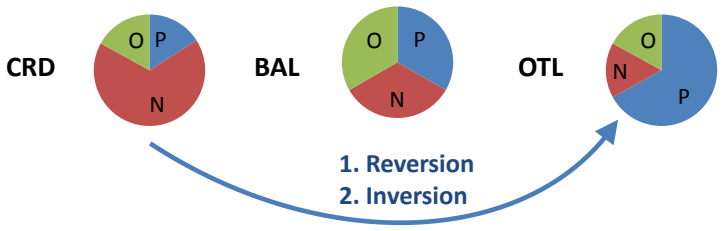

Figure 5.5. Example sentiment biases: We present two approaches to infer the Outlier (OTL) bias from the Crowd (CRD) bias: via reversion of the distribution and via inversion.

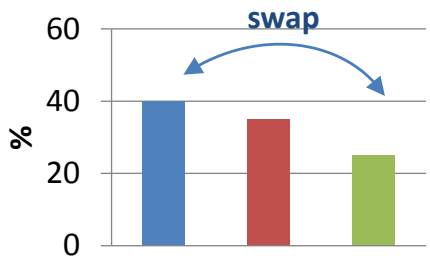

Figure 5.6. Reverting the distribution involves a weight swap between the minority and majority sentiment to obtain OTL from CRD.

of non-topical query aspects $\omega_i \in asp(Q)$ for each dimension M. In the following, we demonstrate how this distribution is estimated *for any single dimension*:

Given any particular M, let $\omega_1, \ldots, \omega_n$ be ranked in increasing order of their values in this distribution, referred to as 'Query Dimension Aspect Distribution'. Further, let $\beta \in [-1; 1]$ be a dimension-specific parameter indicating the direction of the desired bias such that a negative value closer to -1 favors a diversification against the Query Dimension Aspect Distribution (Aktolga & Allan, 2013), whereas a value closer to 0 indicates equal aspect diversification, and a positive value closer to 1 indicates diversification towards the Query Dimension Aspect Distribution. Then, we can define the following bias function that serves for calculating weight distributions for each query aspect ω_i and a user's weight specification β for this particular dimension:

$$\text{bias}(Q, \omega_i, \beta) = \begin{cases} \beta \cdot TOW(Q, \omega_i) + (1 - \beta) \cdot U(Q, \omega_i) & \text{if } \beta \geq 0 \\ |\beta| \cdot AG(Q, \omega_i) + (1 - |\beta|) \cdot U(Q, \omega_i) & \text{otherwise} \end{cases} \tag{5.7}$$

where TOW yields ω_i's weight based on the Query Dimension Aspect Distribution:

$$TOW(Q, \omega_i) = \frac{asp(Q, \omega_i)}{\sum_{\theta \in asp(Q)} asp(Q, \theta)} \tag{5.8}$$

where $asp(Q, \omega_i)$ is just the number of observations with aspect ω_i for Q, which is normalized across all aspect observations for Q. U yields a uniform distribution across all aspects:

$$U(Q, \omega_i) = \frac{1}{|asp(Q)|} \tag{5.9}$$

And finally, AG reverses the values in the Query Dimension Aspect Distribution such that query aspect ω_i is assigned the weight of aspect ω_{n-i+1}:

$$AG(Q, \omega_i) = TOW(Q, \omega_{n-i+1}) \tag{5.10}$$

Note that within this bias framework the actual query aspect distribution is not changed between TOW and AG: we do a value swap to achieve the desired change, which is demonstrated for the sentiment dimension in Figure 5.6. With U however, the distribution is changed.

We demonstrate this with an example for the sentiment dimension. For some query Q, let $asp(Q, positive) = 40\%, asp(Q, negative) = 35\%$, and $asp(Q, neutral) = 25\%$, which is inferred from a dataset. So $|asp(Q)| = 3$, since the sentiment dimension has 3 aspects. Then, we get the following distributions for some example β parameters:

- if $\beta = 1$ (CRD case), then $positive = 40\%, negative = 35\%$, and $neutral = 25\%$;
- if $\beta = 0$ (BAL case), then $positive = negative = neutral = 33.3\%$;

- if $\beta = -1$ (OTL case), then $positive = 25\%, negative = 35\%$, and $neutral = 40\%$;
- if $\beta = 0.5$, then $positive = 36.65\%, negative = 34.15\%$, and $neutral = 29.15\%$;
- if $\beta = -0.5$, then $positive = 29.15\%, negative = 34.15\%$, and $neutral = 36.65\%$.

This bias framework allows a seamless integration of different biases, allowing the user to explore search results in between the spectrum of 'show me an unbiased view of the results' (corresponding to U), 'show me how people think in general about this query' (corresponding to TOW), and 'show me how minority groups think about this query' (corresponding to AG), to give an example for the sentiment dimension.

5.4.2.2 Inverting the Distribution

We consider an alternative formulation of the general bias framework just presented: in this version, the Query Dimension Aspect Distribution is altered by inverting the given distribution. Figure 5.7 shows an example for the sentiment dimension with starting weights for the Crowd bias. The aim is to infer the new weights for the Outlier bias. This is achieved by inverting the Crowd bias weights as shown in Figure 5.8, followed by renormalization to ensure all weights sum to 100%.

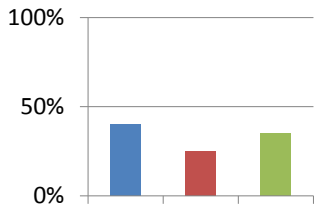

Figure 5.7. Before inverting the distribution with starting weights for the sentiment bias CRD.

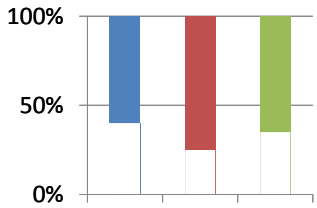

Figure 5.8. After inverting the distribution with new weights for the OTL bias. The weights need to be renormalized before usage.

In the following, we present a formal description of this framework. Let *dist* be a function that yields ω_i's weight based on the "true" Query Dimension Aspect Distribution, similar to TOW in the previous section:

$$dist(Q, \omega_i) = \frac{asp(Q, \omega_i)}{\sum_{\theta \in asp(Q)} asp(Q, \theta)} \tag{5.11}$$

Following the introduction from Section 5.4.2.1.1, the new bias function for calculating weight distributions for each aspect ω_i for a single dimension given query Q and a user's weight specification β then becomes:

$$\text{bias}_{\text{in}}(Q, \omega_i, \beta) = \frac{1}{Z} \cdot (\frac{1+\beta}{2} \cdot dist(Q, \omega_i) + \frac{1-\beta}{2} \cdot (1 - dist(Q, \omega_i))) \tag{5.12}$$

where $Z = \sum_{\psi \in asp(Q)} asp(Q, \psi)$ is the normalization factor – the sum of all the new weights. Normalization is not applied until after the new weights for all aspects ω_i have been calculated, since Z is the sum of those new weights. Note that with this definition, if $\beta = 1$, we get the true Query Dimension Aspect Distribution:

$$\begin{aligned} \text{bias}_{\text{in}}(Q, \omega_i, 1) &= \frac{1}{Z} \cdot dist(Q, \omega_i) \\ &= \frac{1}{Z} \cdot \frac{asp(Q, \omega_i)}{\sum_{\theta \in asp(Q)} asp(Q, \theta)} \end{aligned} \tag{5.13}$$

which is the true distribution normalized with new weights for all query aspects. If $\beta = -1$, we get the inverted true distribution:

$$\text{bias}_{\text{in}}(Q, \omega_i, -1) = \frac{1}{Z} \cdot (1 - dist(Q, \omega_i)) \tag{5.14}$$

And if $\beta = 0$, then this yields

$$\begin{aligned} \text{bias}_{\text{in}}(Q, \omega_i, 0) &= \frac{1}{Z} \cdot (\frac{1}{2} \cdot dist(Q, \omega_i) + \frac{1}{2} \cdot (1 - dist(Q, \omega_i))) \\ &= \frac{1}{2Z} \cdot (\frac{asp(Q, \omega_i)}{\sum_{\theta \in asp(Q)} asp(Q, \theta)} + 1 - \frac{asp(Q, \omega_i)}{\sum_{\theta \in asp(Q)} asp(Q, \theta)}) \\ &= \frac{1}{2Z} \end{aligned} \tag{5.15}$$

So the unnormalized new weight for any query aspect ω_i in the case $\beta = 0$ is always $\frac{1}{2}$. A single dimension has $|asp(Q)|$ number of aspects, so if each one of them has an unnormalized weight of $\frac{1}{2}$, the normalization factor Z becomes $Z = |asp(Q)| \cdot \frac{1}{2}$. So the final normalized weight for each ω_i is

$$\frac{1}{2Z} = \frac{1}{2 \cdot |asp(Q)| \cdot \frac{1}{2}} = \frac{1}{|asp(Q)|} \tag{5.16}$$

which is exactly the equal diversification case U from Section 5.4.2.1.1.

For this approach, too, we show an example for the sentiment dimension. Again, let $asp(Q, positive) = 40\%, asp(Q, negative) = 35\%$, and $asp(Q, neutral) = 25\%$, which is inferred from a dataset for Q. So again $|asp(Q)| = 3$. Then, we get the following distributions for some example β parameters:

- if $\beta = 1$ (CRD case), then $positive = 40\%, negative = 35\%$, and $neutral = 25\%$;
- if $\beta = 0$ (BAL case), then $positive = 33.3\%, negative = 33.3\%$, and $neutral = 33.3\%$;

- if $\beta = -1$ (OTL case), then $positive = 30\%, negative = 32.5\%$, and $neutral = 37.5\%$;
- if $\beta = 0.5$, then $positive = 37.5\%, negative = 34.38\%$, and $neutral = 28.12\%$;
- if $\beta = -0.5$, then $positive = 33\%, negative = 33\%$, and $neutral = 34\%$.

For both approaches to calculating weights in the distribution – whether inverted or reverted – the calculations are done separately for each dimension given a dimension-specific β parameter that indicates how the biases shall be mixed.

5.4.2.2.1 Biases with Sentiments and Time The introduced frameworks above can directly be applied to the sentiment and time dimensions with query sentiment aspects $\sigma \in sent(Q)$ and query time aspects $\pi \in time(Q)$. While the estimation of $U(Q, \omega_i)$ is simply the inverse number of query aspects in each dimension, the other cases need further elaboration. $TOW(Q, \omega_i)$ or $dist(Q, \omega_i)$ employ $asp(Q, \omega)$: for sentiments. In our experiments this is estimated from the fraction of sentiment-tagged documents for Q having the most confident sentiment class $\sigma \in sent(Q)$ as in Chapter 3. For time, the 1:1 mapping of times to documents is less appropriate since we use times mentioned within documents for Q (see Section 5.4.3.3), many of which are often equally relevant. Therefore, we calculate this with the normalized occurrence frequency of π in the pool of relevant documents identified for Q. $AG(Q, \omega_i)$ is based on the TOW case and is therefore handled analogously.

5.4.2.3 Retrieval-Interpolated Diversification

When diversifying search results across multiple dimensions, a decision has to be made about how to manage several dimensions simultaneously. One approach may be to diversify the results list separately from scratch for each dimension, to then subsequently merge those lists. Or we could fine-tune the ranking for each dimension subsequently on the same list: then, further questions would need to be clarified,

such as in which order to apply the reranking for different dimensions. The approach that we choose in this thesis is a direct solution to the desired aims for diversification across multiple dimensions:

1. at each step when choosing the next document to be added to the diversified list, we want to *maximize relevance* not only with respect to certain aspects within one, but within *several* dimensions;

2. at the same time we want to *minimize redundancy* not only among the aspects of one dimension, but within *several* dimensions;

3. the extent to which a certain dimension is emphasized compared to another should be controllable.

To achieve these aims, at each diversification step we choose the document that *best* fulfills the given criteria across several dimensions by linearly interpolating partial scores. The individual dimension scores are merged by weights obtained from a distribution over the dimensions. This allows us to emphasize each dimension precisely as desired.

Algorithm 3 Retrieval Interpolated Diversification Framework.

```
1  S = ∅
2  while|S| < τ and |R| > 0
3        do
4            D* = arg max_{D∈R}  (1 − Σ_M φ_M)RetC(Q) + Σ_M φ_M MC(Q)
5            R = R \ {D*}
6            S = S ∪ {D*}
7  return S
```

Algorithm 3 shows the Retrieval-Interpolated Diversification Framework, which is similar to xQuAD, first introduced by R. L. Santos et al. (2010a) for topical diversity. Here, documents retrieved in R are iteratively added to the new ranked list S. The τ documents are chosen according to the maximization objective function in line 4:

$$D^* = \arg\max_{D \in R} \ (1 - \sum_{M} \varphi_M)\text{RetC}(Q) + \sum_{M} \varphi_M \text{MC}(Q) \tag{5.17}$$

where RetC(Q) is the *retrieval contribution* estimated directly with the retrieval score of D, and MC(Q) are the different *dimension contributions* for the M $\in \mathcal{M}$ dimensions, like sentiment and time contribution, which we will define in two different ways below. The scores from these $\mathcal{M} + 1$ components are interpolated using dimension-specific weights φ_M for diversity estimation. We require that all the interpolation weights sum to 1.0.

5.4.2.3.1 Dimension Contribution by Strength (DCS) In this version of the model we estimate the dimension contribution in the maximization objective function (Equation 5.17) as follows:

$$\text{MC}(Q) = P(D, \bar{S}|Q) \tag{5.18}$$

Here $P(D, \bar{S}|Q)$ measures how much D can contribute to the diversity of S with respect to a particular dimension. Structurally, this resembles xQuAD (R. L. Santos et al., 2010a).

In order to make the model more flexible towards non-topical query aspects scores, we define each document to have a fractional score for each query aspect $\omega \in asp(Q)$. For sentiments, this is straightforward: a document classified as positive with 75% confidence receives a trinary score $P(\sigma = \text{positive}|D) = 0.75$, $P(\sigma = \text{neutral}|D) = 0.25$, and $P(\sigma = \text{negative}|D) = 0$ (Aktolga & Allan, 2013). For time, we had defined in Equation 4.5:

$$P(\pi|D) = \sum_{\kappa \in D} P(\pi|\kappa) \cdot P(\kappa|D)$$
$$= \sum_{\kappa \in D} \frac{|\kappa \cap \pi|}{|\kappa \cup \pi|} \cdot \frac{c(\kappa, D)}{\sum_{\mu \in D} c(\mu, D)} \tag{5.19}$$

where $P(\kappa|D)$ represents the likelihood of time κ occurring in D, estimated through normalized frequency counts for those time mentions. $P(\pi|\kappa)$ expresses how well κ covers π. Given two time intervals, π and κ, $P(\pi|\kappa)$ determines whether they overlap and how significant this overlap is.

Given this information, we can decompose $P(D, \bar{S}|Q)$ as in Chapter 3:

$$P(D, \bar{S}|Q) \stackrel{rank}{=} \sum_{\omega \in asp(Q)} P(D|\omega) \cdot P(\bar{S}|\omega) \cdot P(\omega|Q) \tag{5.20}$$

where $P(\bar{S}|\omega)$ denotes the likelihood of ω not being satisfied by the documents already chosen into S, and $P(\omega|Q)$ stands for the importance of non-topical query aspect ω to query Q, which was explained in detail in Section 5.4.2.1. For practical purposes, in the experiments we estimate $P(D|\omega) \stackrel{rank}{=} P(\omega|D)$ by applying Bayes' Rule and omitting the constants. For our two dimensions, sentiment and time, $P(\omega|D)$ translates into the sentiment score of the document $P(\sigma|D)$ and the time score $P(\pi|D)$ respectively. Analogous to Chapter 3, the final derivation of Equation 5.20 is:

$$P(D, \bar{S}|Q) \stackrel{rank}{=} \sum_{\omega \in asp(Q)} P(D|\omega) \cdot P(\omega|Q) \cdot \prod_{D_j \in S} 1 - P(D_j|\omega) \tag{5.21}$$

Thus, Equation 5.21 estimates the diversity of D by considering how well it represents each non-topical query aspect, which is weighted by how important that query aspect is to Q. This whole part is demoted depending on how strong the documents already chosen into S are with respect to that query aspect.

5.4.2.3.2 Dimension Contribution by Strength and Frequency (DCSF)

Alternatively, we can estimate the dimension contribution part in Equation 5.17 as follows:

$$\text{MC}(Q) = P(D|Q) \cdot (1 - P(S|Q)) \tag{5.22}$$

Following the same step-by-step derivation as in Chapter 3, we obtain:

$$P(D|Q) \cdot (1 - P(S|Q)) = \sum_{\omega \in asp(Q)} P(D|\omega) \cdot P(\omega|Q) \cdot P(\bar{\omega}|S) \tag{5.23}$$

where $P(\bar{\omega}|S)$ is the likelihood of S not having non-topical query aspect ω. We can define this in a dimension-specific way: for sentiments $P(\bar{\omega}|S) \Rightarrow P(\bar{\sigma}|S) = 1 - \frac{sent(\sigma,S)}{|S|}$, which is the fraction of documents in S not having dominant sentiment σ, whereas for time we defined in Equation 4.6:

$$P(\bar{\omega}|S) \Rightarrow P(\bar{\pi}|S) = \frac{time(\bar{\pi}, S)}{|S|} \tag{5.24}$$

which is the fraction of documents in S not covering time π to at least 50%, which was tuned on a held-out dataset during experimentation. For estimating coverage between times $\kappa \in D \in S$ and π we use $\frac{|\kappa \cap \pi|}{|\kappa \cup \pi|}$ as in Equation 5.19. We set $P(\omega|S) = 0$ if $S = \emptyset$ to avoid zero division in the first iteration.

Instead of considering the strength of sentiment or time scores, with this alternative formulation the *frequency* of documents in S with certain dominant non-topical query aspects is directly used to control diversity.

5.4.2.4 Diversity by Proportionality

As another diversification framework we consider PM-2 (Algorithm 4), which is based on the Sante-Laguë method for seat allocation and is adapted here for diversification with multiple dimensions and non-topical aspects. In prior work this model is applied to the sentiment dimension only (Aktolga & Allan, 2013) and to topical

Algorithm 4 Diversity by Proportionality (PM-2).

1 $S = \emptyset$
2 $\forall M \;\; \forall \omega \;\; s_\omega = 0$
3 **while**$|S| < \tau$ and $|R| > 0$
4 **do**
5 **for** $M \in \mathcal{M}$
6 **do**
7 **for** $\omega \in asp(Q)$
8 **do**
9 $quotient[\omega] = \frac{v_\omega}{2s_\omega + 1}$
10 $\omega^* = \arg\max_\omega quotient[\omega]$
11 $D^* = \arg\max_{D \in R} \;\; \sum_M \varphi_M \cdot [\lambda \cdot quotient[\omega^*] \cdot P(D|\omega^*) + (1 - \lambda) \sum_{\omega \neq \omega^*} quotient[\omega] \cdot P(D|\omega)]$
12 $R = R \setminus \{D^*\}$
13 $S = S \cup \{D^*\}$
14 **for** $M \in \mathcal{M}$
15 **do**
16 **for** $\omega \in asp(Q)$
17 **do**
18 $s_\omega = s_\omega + \frac{P(D^*|\omega)}{\sum_{\gamma \in asp(T)} P(D^*|\gamma)}$
19 **return** S

aspects (Dang & Croft, 2012). In Section 3.2.2.3, the algorithm is explained in detail for the sentiment dimension. Here we describe modified components for multiple dimensions.

The initialization of the variables is similar to Algorithm 2, except for the fact that now we have each one s_ω and v_ω for each non-topical query aspect $\omega \in asp(Q)$ of each dimension $M \in \mathcal{M}$. These dimension-specific variables still have the same role: v_ω indicates the number of relevant documents aspect ω should have, whereas s_ω represents the number of documents actually present in the list for ω.

In each iteration of the while loop in lines 3-18 the query aspect-specific quotient is calculated for each dimension M and its non-topical query aspects (lines 5-9, Algorithm 4). Then the aspect(s) to focus on in the current iteration are chosen (line 10), one for each dimension. The next document for S is determined by considering its relevance to the chosen query aspect ω^* versus its relevance to all the other query aspects within that dimension. We combine this across multiple dimensions M via interpolation with weights φ_M, which constitute a probability distribution over the dimensions. To enforce this, we require that all dimension interpolation weights

sum to 1.0, i.e., $\sum_{M \in \mathcal{M}} \varphi_M = 1.0$. Instead of letting λ be a common interpolation parameter, we set:

$$\lambda = \frac{P(D|\omega^*)}{P(D|\omega^*) + \rho} \tag{5.25}$$

to achieve Dirichlet-like smoothing with parameter ρ tuned in the experiments. $P(D|\omega^*)$ is the dimension-specific relevance score of the document as detailed in Section 5.4.2.3.1: for sentiments, this is the sentiment score and for time we use the fractional time score described in Equation 5.19.

5.4.2.4.1 Diversity by Proportionality with Minimum Available Votes (PM-2M) Following Chapter 3, we adapt this modification to PM-2 to yield PM-2M:

$$quotient[\omega] = \frac{\min(v_\omega, l_\omega)}{2s_\omega + 1} \tag{5.26}$$

which is a modified calculation of the quotient to avoid the exploitation of aspects in early ranks for which limited data is available in the retrieved list. This can happen if a particular aspect is underrepresented in the top K documents retrieved from a search system, resulting in suboptimal diversification. With this modification we replace an overestimating v_ω with l_ω, the actual amount of documents available with aspect ω, to achieve better diversification.

5.4.3 Experimental Setup

5.4.3.1 Data

Retrieval Corpus We use the TREC Blog Track data 2006-2008 (Ounis et al., 2006) as retrieval corpus for all our experiments. For preparation, the DiffPost algorithm is applied for better retrieval as shown in prior work (Lee et al., 2008; Nam et al., 2009). Further, we perform stop word removal and Porter stemming.

Queries and Retrieval Model We split the 150 TREC Blog Track 2008 queries into 3 non-overlapping randomly chosen sets of size 50 each in order not to bias training or testing towards a specific year: split 1 is used for training and tuning parameters; the results in this work are reported on split 2, and split 3 is reserved for sentiment classifier training. For our diversification experiments, we use a strong retrieval baseline: the queries' stopped title and description texts are combined for use with the Sequential Dependence Model in Lemur/Indri (Metzler & Croft, 2005), smoothed using Dirichlet ($\mu = 10,000$). All diversification models are applied to the top $K = 50$ retrieved documents as determined during training. The retrieval scores are normalized to yield document likelihood scores.

5.4.3.2 Sentiments

Sentiment Classification The sentiment classifier is trained as a logistic regression model using Liblinear (Fan et al., 2008) with default settings as in Chapter 3. For this, we utilize the judged documents from the 50 split 3 TREC Blog Track queries. Training is done for three classes – positive, negative, and neutral to obtain probability estimates that are employed as fractional scores for sentiment estimation. As features we extract Sentiwordnet 3.0 terms with their length-normalized term frequencies in the documents (Baccianella et al., 2010).

Query Sentiment Aspects Distribution Estimation Given a query, its sentiment aspects distribution is estimated in the form of opinion relevance judgments from the TREC 2008 Blog Track (Ounis et al., 2006) following Chapter 3. To observe diversification performance at various sentiment classification accuracies during the experiments (Section 5.4.4.1), classification labels are simulated as described in Chapter 3. During the evaluation, we use the full set of relevance judgments.

5.4.3.3 Time

Extracting Times From Documents This is handled in the same way as described in Section 4.4.2.2.

Truth Judgments The TREC 2008 Blog Track judgments already include sentiment-level judgments but there are no time-specific judgments. We augment these with time judgments as described in Section 4.4.2.2.

Query Time Aspects Estimation This is handled in the same way as described in Section 4.4.2.2.

5.4.3.4 Biases

Although the general bias framework allows arbitrary values of β to specify desired diversity, evaluation is clearer with a limited number of possibilities. We thus select the two endpoints and the mid-point in each dimension: equal diversification with $\beta = 0$ in Equation 5.7, diversification towards the query dimension aspects distribution with $\beta = 1$, and diversification against the query dimension aspects distribution with $\beta = -1$. For sentiments, we refer to the biases as 'Balance' (BAL), 'Crowd' (CRD), and 'Outlier' (OTL) as in Chapter 3, whereas for time we use 'Equal' (EQ), 'Spike' (SPK), and 'Slab' (SLB) as in Chapter 4. For all experiments we use the 'Reverting the Distribution' variation of the bias framework in the experiments (Section 5.4.2.1.1), except for Section 5.4.4.4, where we show some results with $\beta = -1$ for both dimensions with the 'Inverting the Distribution' version of the framework (Section 5.4.2.2).

5.4.3.5 Evaluation Measures

The goal of the experiments in the next section is to see how well the different biases provide diverse results for the user to consider. To evaluate diversity, we use standard evaluation measures that were designed for topical diversity: Precision-

IA (Agrawal et al., 2009), s-recall (Zhai et al., 2003), α-NDCG (C. L. Clarke et al., 2008), ERR-IA (Ashkan & Clarke, 2011), and NRBP (C. L. Clarke et al., 2009). In order to measure non-topical diversity with a chosen bias, we implement all the measures in their intent-aware version (Agrawal et al., 2009; Ashkan & Clarke, 2011). Measure-IA for a query Q combines the scores for multiple dimensions $M \in \mathcal{M}$ as follows:

$$\text{measure-IA}(Q) = \sum_{M} \phi_M \cdot \sum_{\omega} P(\omega|Q) \cdot \text{measure}(Q|\omega) \tag{5.27}$$

where $P(\omega|Q)$ defines the weight for the dimension-aspect-specific result yielded by measure$(Q|\omega)$ and ϕ_M is the dimension-specific weight balancing the scores for different dimensions. For sentiments, we estimate ϕ_{sent} from the provocativeness of Q's topic (Cartright et al., 2009):

$$\phi_{sent} = PROV(Q(T)) = \frac{\sum_{t \in rel(T)} \text{subjectivity}(t)}{|rel(T)|} \tag{5.28}$$

which is the fraction of subjective documents for T. This can be estimated with positive, negative, and mixed judgments for T. For time, we set $\phi_{time} = 1 - \phi_{sent}$. This way, the more provocative the topic is, the more important the sentiment dimension is in Equation 5.27, and the weight for the time dimension is adjusted accordingly. Alternatively we tried $\phi_{sent} = \phi_{time} = 0.5$, which yields similar results.

5.4.4 Results

In this section we discuss the results of the retrieval baseline SDM and the diversification models proposed in Section 5.4.2, DCS, DCSF, PM-2, and PM-2M with various biases for the sentiment and time dimensions. We have several interpolation parameters tuned on the train split: φ_M is initialized with φ_{sent} and φ_{time}, each of which is tuned in 0.1 steps separately for each model and bias. For the proportionality models we tune $\rho \in [5, 20, 500, 5000]$ similarly. The results are presented with

fixed parameters on the test split, and the evaluation is performed with the time-augmented TREC 2008 Blog Track judgments at rank 20 (see Section 5.4.3.3). For α-NDCG, ERR-IA, and NRBP we set $\alpha = \beta = 0.5$. Statistical significance tests are reported using the paired two-sided t-test with p-value < 0.05; smaller p-values are explicitly stated with the results.

5.4.4.1 Straight-Bias Experiments

		Time		
	Biases	SPK	EQ	SLB
Sentiment	CRD	Base	–	Add
	BAL	Add	Base	–
	OTL	Add	–	Base

Table 5.2. Nine possible bias combinations for our two dimensions, highlighting base and additional bias combinations we use in the experiments. Omitted cases are marked as '–.'

In this style of experiments we use the **same** bias combination during diversification and evaluation, although the sources of bias estimation differ, as detailed in Section 5.4.3. For example, diversification and evaluation is done with Crowd + Spike, but the bias information for the experiments comes from our sentiment classifier and from Wikipedia versus truth relevance judgments in the evaluation. Since we have two dimensions with three extreme target biases each (Section 5.4.2.1 and Section 5.4.3.4), a total of nine bias combinations are possible for diversification, as shown in Table 5.2. Bias combinations using the same β for both dimensions are marked as 'Base', whereas those using different β's are denoted as 'Add' (for 'additional'). We include all base bias cases and some additional cases in the analysis. Three additional cases involving BAL or EQ as one of the components are omitted.

In Figure 5.9 we observe the results across three measures with varying sentiment classifier accuracies on the x-axis. Note that we hold query time aspect estimations

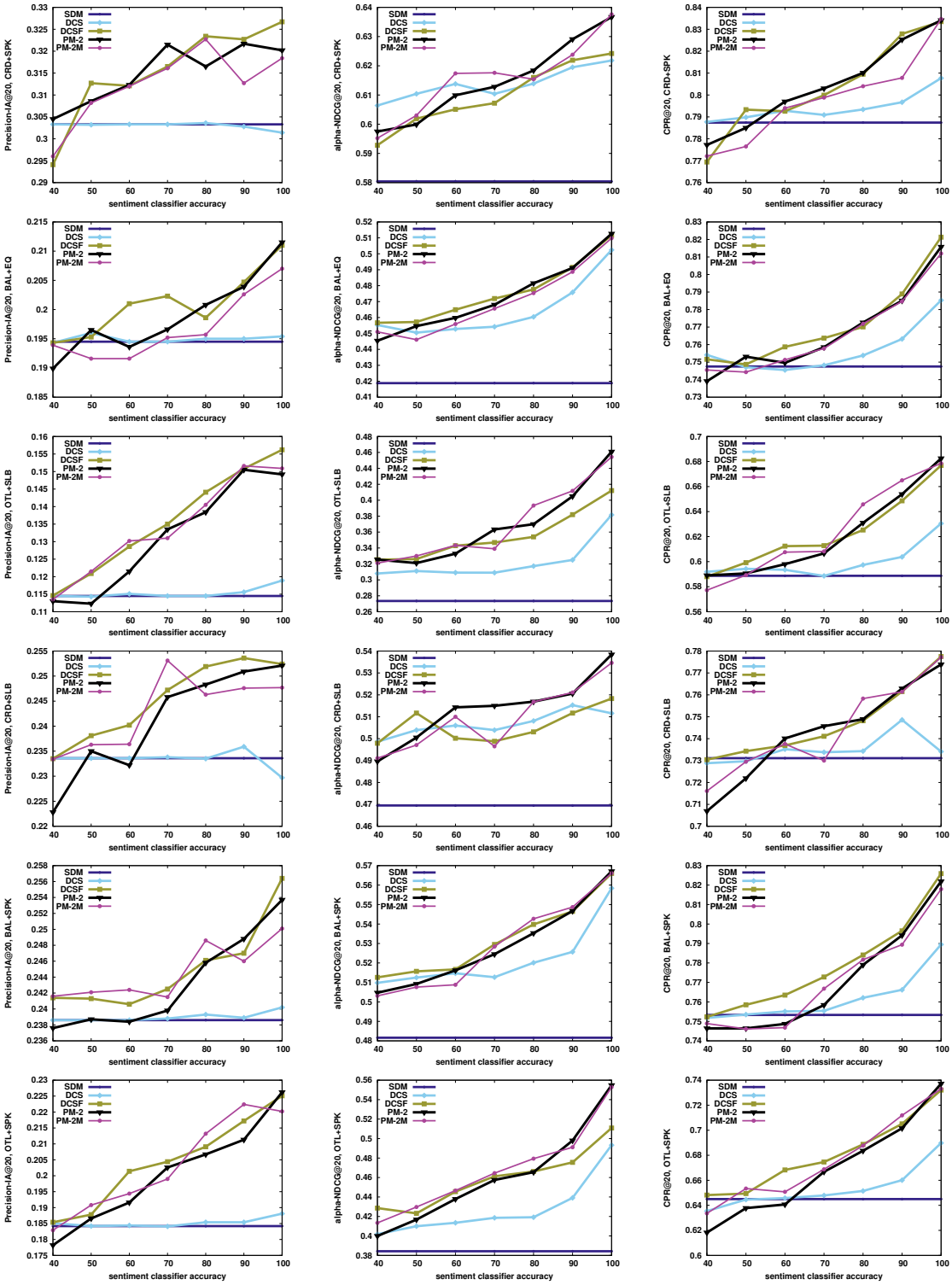

Figure 5.9. Straight-Bias Experiments: varying sentiment classifier accuracies on the x-axis and each one measure and bias on the y-axis.

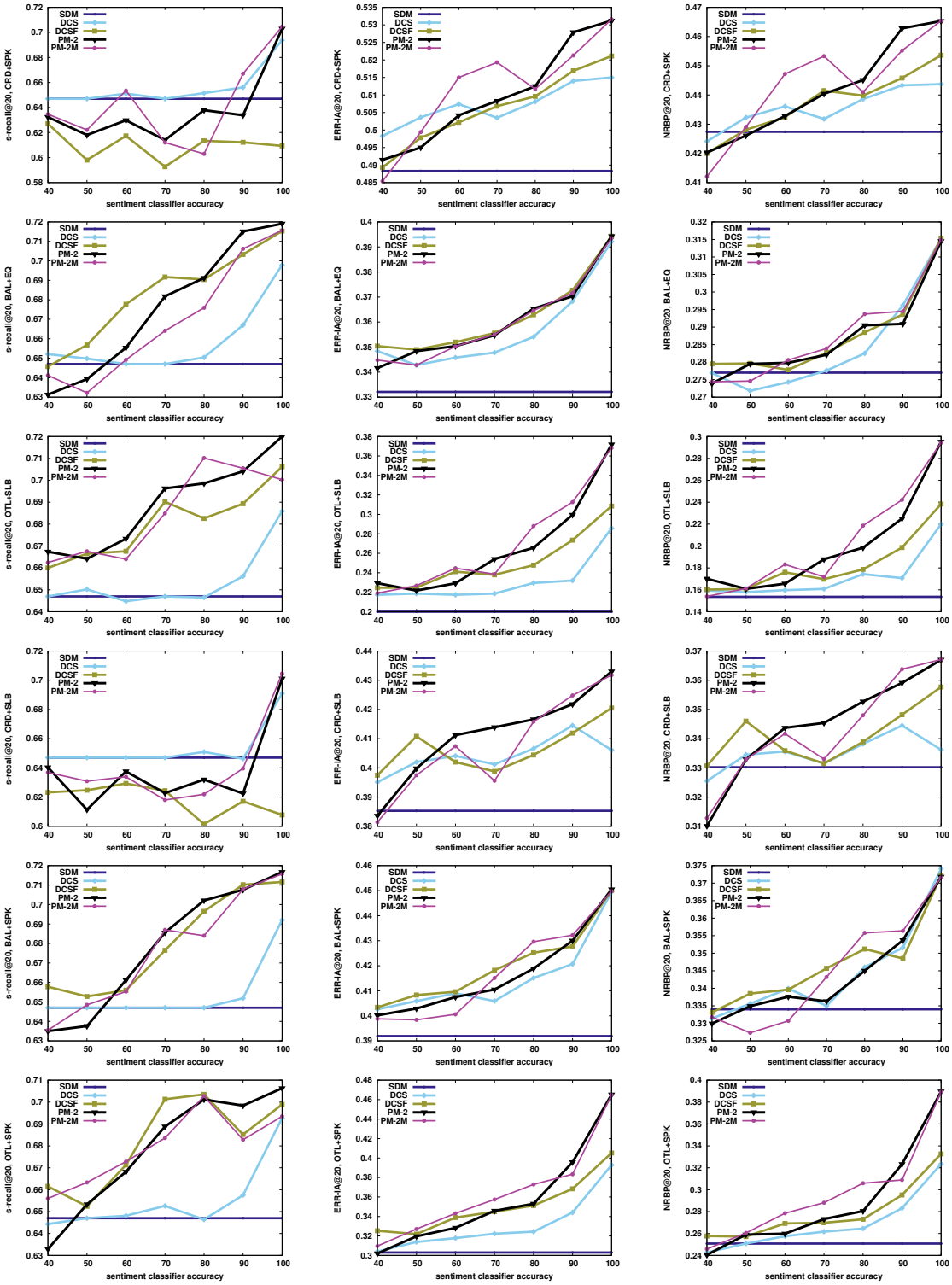

Figure 5.10. Straight-Bias Experiments: varying sentiment classifier accuracies on the x-axis and each one measure and bias on the y-axis.

constant in this experiment while varying sentiment classification. In all graphs the results improve for all methods with better classification. When sentiments are perfect, DCSF, PM-2 and PM-2M perform significantly better (p-value < 0.05) than the SDM baseline in all graphs, and also over the DCS approach in almost all graphs. These three models seem most stable in all bias settings. The main cutoff point for significance is around the 60-70% sentiment classification marker for the CRD+SPK, OTL+SLB, and OTL+SPK biases according to Precision-IA@20 and CPR@20, however α-NDCG@20 can tolerate much lower accuracies. For other biases, the borderline for significance is much higher: 80-90% for BAL+EQ and 70-80% for CRD+SLB and BAL + SPK. Although a 1:1 comparison between these results and those with single dimension sentiment diversification in Chapter 3 is not applicable because the evaluations are different (judgments for two dimensions instead of one), we can observe that with the addition of time, achieving significant results is much harder. Both for Precision-IA@20 and CPR@20 the borderline for significance is much higher here, whereas α-NDCG@20 seems unaffected by the inclusion of an additional dimension. In Figure 5.10 we have another set of straight-bias results with three other measures, s-recall@20, ERR-IA@20, and NRBP. The results for s-recall@20 are most interesting: whenever the Crowd bias is involved, such as with CRD+SPK or CRD+SLB, the results for most approaches become worse than the SDM baseline if the sentiment labels are not perfect. We observed a similar trend in Chapter 3 for the Crowd bias. For Balance bias combinations, such as BAL+EQ and BAL+SPK there are also similarities with results in Chapter 3: DCSF and the proportionality models remain significant over the SDM baseline until 60-70% classification accuracy. The ERR-IA@20 results also seem to be dependent on the type of sentiment bias: for results with Balance, such as BAL+EQ and BAL+SPK, the results remain significant until lower classification accuracies, however when the Outlier bias is used such as for OTL+SPK and OTL+SLB, at least 70% accuracy is required for the results

to be significantly better than the SDM baseline. NRBP on the other hand cannot tolerate bad classification labels, and almost all approaches perform worse than the SDM baseline with 40%-50% classification accuracy. In Section 5.4.4.3 we compare the results for the three base bias cases to perfect classification of query sentiment *and* time aspects.

5.4.4.2 Cross-Bias Experiments

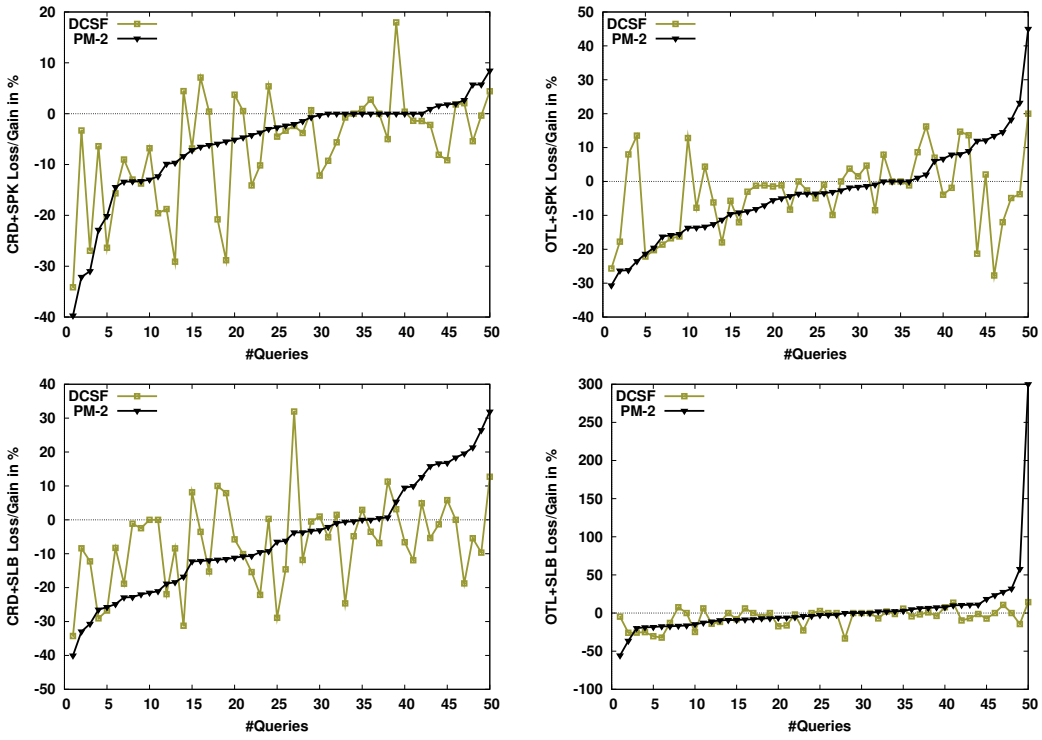

Figure 5.11. Cross-Bias Experiments: Relative loss/gain when diversifying with BAL+EQ and evaluating with different bias combinations.

If we reflect on the general bias framework introduced in Section 5.4.2.1, only the equal diversification case with $\beta = 0$ does not require a dataset for bias estimation. Therefore, this is typically the default bias to be employed for all dimensions if no reliable dataset is available. So if equal diversification is used for the time and sentiment dimensions due to a lack of a suitable dataset, how much performance is lost or gained in this case? To answer this, we diversify with BAL+EQ, and evaluate

Measure	**Precision-IA@20**		**α-NDCG@20**		**CPR@20**	
Approach	**DCSF**	**PM-2**	**DCSF**	**PM-2**	**DCSF**	**PM-2**
CRD+SPK, straight	**0.3267**	**0.3202**	0.6242	**0.6367**	**0.8335**	**0.8341**
CRD+SPK, cross	0.2989	0.2984	0.6250	0.6204	0.8208	0.8122
OTL+SLB, straight	**0.1562**	0.1492	**0.4121**	**0.4609**	0.6769	0.6825
OTL+SLB, cross	0.1435	0.1435	0.3913	0.3967	0.6700	0.6673
OTL+SPK, straight	**0.2251**	**0.2262**	**0.5109**	**0.5548**	**0.7321**	0.7372
OTL+SPK, cross	0.2146	0.2134	0.4909	0.4966	0.7259	0.7216
CRD+SLB, straight	**0.2524**	**0.2521**	0.5183	0.5384	**0.7775**	**0.7739**
CRD+SLB, cross	0.2267	0.2282	0.5258	0.5221	0.7646	0.7578

Table 5.3. Average Cross-Bias Experiment results for three measures with DCSF and PM-2. All cross-runs are diversified based on BAL+EQ. Bold straight results are significantly better than their cross-bias counterparts (p-value < 0.02).

this result with four bias combinations: CRD+SPK, OTL+SLB, OTL+SPK, and CRD+SLB. This performance is compared to using the actual intended bias (such as CRD+SPK) during diversification **and** evaluation. We note the performance loss or gain on a query-by-query basis on our test queries in Figure 5.11 for DCSF and PM-2 with Precision-IA@20. The results are sorted according to PM-2 along the x-axis: in all four graphs we can see either no change or a loss in performance for around 40 queries by using BAL+EQ, and a gain for about 10 queries. While query-by-query results strongly vary between DCSF and PM-2, the number of queries for which we observe gains/losses is similar. It is interesting to note that the OTL+SLB graph shows the least variation, whereas OTL+SPK is more variable similar to the other two bias combinations. In the OTL+SLB graph we have one topic with 300% improvement for PM-2: this seems not to be an interesting outlier due to very low precision that jumped from 0.008 to 0.03. Overall, the results are very similar across other measures.

With averaged results in Table 5.3, we observe the largest losses for CRD+SLB with 10.2% for DCSF and even a small gain of 1.4% for CRD+SLB. For PM-2, the maximum loss is 13.9% with OTL+SLB. On average, across all biases, around 3.7%

significant performance is lost (p-value < 0.02) by using the BAL+EQ biases for DCSF and 5.4% for PM-2.

To conclude, although BAL+EQ is a reasonable alternative, the actual intended bias should be inferred from a suitable dataset to get maximum diversification performance.

5.4.4.3 Perfect Query Sentiment and Time Aspects

Measure	**Precision-IA@20**					
Bias	**CRD+SPK**	**BAL+EQ**	**OTL+SLB**	**BAL+SPK**	**OTL+SPK**	**CRD+SLB**
SDM baseline	0.3033	0.1945	0.1145	0.2386	0.1842	0.2336
DCS	**0.3296**	**0.2203**	**0.1424**	**0.2632**	**0.2067**	**0.2489**
DCSF	**0.3507***	**0.2211**	**0.1706***	**0.2783***	**0.2468***	**0.2704***
PM-2	**0.3470***	**0.2233**	**0.1687***	**0.2786***	**0.2444***	**0.2642***
PM-2M	**0.3398***	**0.2070**	**0.1633***	**0.2703**	**0.2419***	**0.2616***
Measure	**α-NDCG@20**					
Bias	**CRD+SPK**	**BAL+EQ**	**OTL+SLB**	**BAL+SPK**	**OTL+SPK**	**CRD+SLB**
SDM baseline	0.5804	0.4188	0.2735	0.4816	0.3843	0.4695
DCS	**0.6425**	**0.5229**	**0.4418**	**0.5763**	**0.5093**	**0.5623**
DCSF	**0.6290**	**0.5256**	**0.4436**	**0.5789**	**0.5218**	**0.5660**
PM-2	**0.6341**	**0.5168**	**0.4939***	**0.5683**	**0.5518***	**0.5827**
PM-2M	0.6095	**0.5098**	**0.4575**	**0.5521**	**0.5458***	**0.5389**
Measure	**CPR@20**					
Bias	**CRD+SPK**	**BAL+EQ**	**OTL+SLB**	**BAL+SPK**	**OTL+SPK**	**CRD+SLB**
SDM baseline	0.7874	0.7475	0.5887	0.7533	0.6450	0.7311
DCS	**0.8253**	**0.8126**	**0.6742**	**0.8160**	**0.7105**	**0.7759**
DCSF	**0.8490***	**0.8397***	**0.7068***	**0.8448***	**0.7484***	**0.8096***
PM-2	**0.8447***	**0.8272**	**0.7065***	**0.8390***	**0.7503***	**0.8114***
PM-2M	**0.8262**	**0.8120**	**0.6885**	**0.8197**	**0.7426***	**0.7903**

Table 5.4. 'Perfect' results for three measures. Bold entries are significantly better than the SDM baseline (p-value < 0.02), whereas bold and starred entries yield a significant gain over DCS (p-value < 0.04).

In the next set of experiments in Tables 5.4 and 5.5 we show the maximum diversification performance possible under ideal circumstances. For these results we utilize *relevance judgments* for Query Aspects Distribution estimation during diversification *and* evaluation for *both* dimensions, thus yielding perfect "oracle" query aspects and bias estimation. Comparing the results to the graphs in Figure 5.9, all approaches unsurprisingly perform better with perfect data. For BAL+EQ in particular, here DCS yields much better results than in Figure 5.9, significantly improving the results over the SDM baseline like the other approaches. This is also the case with α-NDCG

Measure	s-recall@20					
Bias	**CRD+SPK**	**BAL+EQ**	**OTL+SLB**	**BAL+SPK**	**OTL+SPK**	**CRD+SLB**
SDM baseline	0.6470	0.6470	0.6470	0.6470	0.6470	0.6470
DCS	**0.7280**	**0.7294**	**0.7137**	**0.7112**	**0.7056**	**0.7087**
DCSF	0.5995	**0.7071**	**0.7048**	**0.7014**	0.6937	0.6416
PM-2	0.6739	**0.7242**	**0.7050**	**0.6986**	0.6810	**0.7094**
PM-2M	0.6605	**0.7157**	**0.7052**	0.6877	0.6703	**0.7089**
Measure	**ERR-IA@20**					
Bias	**CRD+SPK**	**BAL+EQ**	**OTL+SLB**	**BAL+SPK**	**OTL+SPK**	**CRD+SLB**
SDM baseline	0.4883	0.3321	0.2000	0.3919	0.3030	0.3853
DCS	**0.5306**	**0.4060**	**0.3359**	**0.4637**	**0.4050**	**0.4556**
DCSF	**0.5278**	**0.4104**	**0.3372**	**0.4668**	**0.4187**	**0.4617**
PM-2	**0.5283**	**0.3950**	**0.4078***	**0.4519**	**0.4616***	**0.4809**
PM-2M	0.5048	**0.3935**	**0.3708**	**0.4381**	**0.4575***	**0.4348**
Measure	**NRBP**					
Bias	**CRD+SPK**	**BAL+EQ**	**OTL+SLB**	**BAL+SPK**	**OTL+SPK**	**CRD+SLB**
SDM baseline	0.4274	0.2770	0.1536	0.3340	0.2508	0.3302
DCS	**0.4576**	**0.3282**	**0.2621**	**0.3886**	**0.3361**	**0.3816**
DCSF	**0.4590**	**0.3297**	**0.2654**	**0.3905**	**0.3486**	**0.3922**
PM-2	**0.4621**	**0.3136**	**0.3291***	**0.3744**	**0.3876**	**0.4109**
PM-2M	0.4397	**0.3149**	**0.2964**	**0.3626**	**0.3853**	**0.3674**

Table 5.5. 'Perfect' results for three measures. Bold entries are significantly better than the SDM baseline (p-value < 0.02), whereas bold and starred entries yield a significant gain over DCS (p-value < 0.04).

for CRD+SPK and BAL+SPK, where DCS even outperforms the other methods with perfect information. However, since this method is not so stable in our straight-bias experiments, it proves rather sensitive to noisy bias and query aspects estimation.

5.4.4.4 Perfect Query Sentiment and Time Aspects with Inverted Distribution

In Table 5.6 we also take a look at diversifying with perfect data when the Query Aspects Distribution for the Outlier or Slab biases is calculated by inverting the distribution, as described in Section 5.4.2.2. Note that we utilize the inversion approach during diversification *and* evaluation. First of all we note that the results are slightly better than the ones presented in Tables 5.4 and 5.5, which were calculated by reverting the distribution for Outlier and Slab biases during diversification and evaluation. This means that the diversification methods can deal better with inverted distributions. While significance results are similar for both approaches to handling the distributions, however when inverting the distribution, the DCSF and

Bias	OTL+SLB					
Measures	**Precision-IA@20**	**s-recall@20**	**α-NDCG@20**	**ERR-IA@20**	**CPR@20**	**NRBP**
SDM baseline	0.1513	0.6470	0.3518	0.2682	0.6953	0.2159
DCS	**0.1754**	**0.7137**	**0.4738**	**0.3602**	**0.7656**	**0.2830**
DCSF	**0.1937***	**0.7226**	**0.4890**	**0.3733**	**0.8004***	**0.2948**
PM-2	**0.1933***	**0.7248**	**0.4822**	**0.3634**	**0.7915***	**0.2834**
PM-2M	**0.1894***	**0.7116**	**0.4703**	**0.3562**	**0.7831**	**0.2789**
Bias	OTL+SPK					
Measures	**Precision-IA@20**	**s-recall@20**	**α-NDCG@20**	**ERR-IA@20**	**CPR@20**	**NRBP**
SDM baseline	0.2006	0.6470	0.4218	0.3354	0.7041	0.2803
DCS	**0.2335**	**0.7274**	**0.5412**	**0.4293**	**0.7761**	**0.3545**
DCSF	**0.2531***	**0.7006**	**0.5419**	**0.4320**	**0.8051***	**0.3565**
PM-2	**0.2529***	**0.6966**	**0.5328**	**0.4186**	**0.8031***	**0.3432**
PM-2M	**0.2474***	0.6854	**0.5251**	**0.4137**	**0.7916**	**0.3397**
Bias	CRD+SLB					
Measures	**Precision-IA@20**	**s-recall@20**	**α-NDCG@20**	**ERR-IA@20**	**CPR@20**	**NRBP**
SDM baseline	0.2540	0.6470	0.5103	0.4210	0.7786	0.3630
DCS	**0.2782**	**0.7256**	**0.5823**	**0.4667**	**0.8204**	**0.3901**
DCSF	**0.2904***	0.6261	**0.5758**	**0.4685**	**0.8418***	**0.3958**
PM-2	**0.2890***	**0.7120**	**0.5897**	**0.4786**	**0.8470***	**0.4076**
PM-2M	**0.2685**	**0.7046**	**0.5738**	**0.4666**	**0.8255**	**0.3988**

Table 5.6. 'Perfect' results with inverted distribution for OTL and SLB for all measures with three bias combinations. Bold entries are significantly better than the SDM baseline (p-value < 0.002), whereas bold and starred entries yield a significant gain over DCS (p-value < 0.04).

proportionality models do not perform significantly better than DCS according to the CRD+SLB and OTL+SPK biases to the extent that they do so when reverting the distribution for those bias combinations.

5.4.4.5 Collapsing Dates for Query Time Aspects

Analogous to Section 4.4.3.2, we conduct experiments with collapsed query times and weights for the time dimension. We observe similar results: collapsing dates only helps the results for the Slab bias, such as CRD+SLB and OTL+SLB biases. The results for the former bias combination are shown in Table 5.8 and the ones for the latter one in Table 5.7. Over all measures and approaches in Table 5.7, we note significant improvements with p-value < 0.02, with an overall average improvement of 7.5%. In Table 5.8 we have a lower p-value < 0.009 with the same significance levels over all approaches. To have a counter example, we show the results for CRD+SPK in Table 5.9. Here, we note significant losses in all but one case (p-value < 0.02). Why does collapsing dates only help the Slab bias? We can give the same argument

as in Section 4.4.3.2: the Slab bias emphasizes the tail distribution of times for the query. However, since the tail is usually very flat and sparse, important small-interval time ranges are more strongly emphasized by collapsing. This does not help the Spike bias though, since for most topics the front of the distribution has times with larger intervals under that bias, so those do not heavily profit from collapsing times and weights.

OTL+SLB	**Precision-IA@20**	**Improvement**
SDM baseline	0.1292	+12.8%
DCS	0.1348	+13.4%
DCSF	0.1727	+10.6%
PM-2	0.1658	+11.1%
PM-2M	0.1673	+10.9%
OTL+SLB	**α-NDCG@20**	**Improvement**
SDM baseline	0.2981	+9.0%
DCS	0.4184	+9.6%
DCSF	0.4365	+5.9%
PM-2	0.4792	+4.0%
PM-2M	0.4759	+4.7%
OTL+SLB	**CPR@20**	**Improvement**
SDM baseline	0.6136	+4.2%
DCS	0.6647	+5.4%
DCSF	0.7036	+3.9%
PM-2	0.7076	+3.7%
PM-2M	0.7082	+4.4%

Table 5.7. Results with collapsed dates for OTL+SLB with relative improvements with respect to not collapsing dates. *All* entries are significantly better than their counterpart non-collapsed results (p-value < 0.02).

5.4.4.6 Temporally Unambiguous Queries

In several parts of this thesis (Chapter 2, Section 4.1, and Section 5.1) we pointed out that there are two broad classes of temporal queries: the temporally ambiguous ones, and the temporally unambiguous ones. The former refers to queries that have several relevant time points/intervals, whereas the latter in general only has one relevant specified time point/interval. We also stated that in terms of diversification,

CRD+SLB	Precision-IA@20	Improvement
SDM baseline	0.2482	+6.3%
DCS	0.2510	+9.3%
DCSF	0.2703	+7.1%
PM-2	0.2639	+4.7%
PM-2M	0.2652	+7.1%
CRD+SLB	**α-NDCG@20**	**Improvement**
SDM baseline	0.4941	+5.2%
DCS	0.5470	+6.9%
DCSF	0.5456	+5.3%
PM-2	0.5629	+4.6%
PM-2M	0.5567	+4.1%
CRD+SLB	**CPR@20**	**Improvement**
SDM baseline	0.7560	+3.4 %
DCS	0.7760	+5.7 %
DCSF	0.8022	+3.2 %
PM-2	0.8032	+3.8 %
PM-2M	0.8048	+3.6%

Table 5.8. Results with collapsed dates for CRD+SLB with relative improvements with respect to not collapsing dates. *All* entries are significantly better than their counterpart non-collapsed results (p-value < 0.009).

in general, temporally ambiguous queries may be more suitable, but that temporally unambiguous queries should not be excluded from the experiments. In this section, we now want to explicitly look at a few temporally unambiguous queries/topics that we have in our training and test sets to see whether there are any differences between them standing out from the average results for all queries.

First, we look at a subset of the relevance judgments for a few temporally unambiguous queries in Table 5.10: for each topic we list one relevant unambiguous time, which we consider as most important. In addition, we also list some other relevant time intervals at which for instance follow-up events happened, which are related to the main one. For example, the first topic number '851' is about a 2005 movie, March of the Penguins. Undoubtedly, 2005 is the most important time for this topic. However, there are many other times at which important events happened: for instance,

CRD+SPK	Precision-IA@20	Loss
SDM baseline	0.2800	-7.7%
DCS	0.2787	-7.5%
DCSF	0.3043	-6.9%
PM-2	0.2982	-6.9%
PM-2M	0.2957	-7.1%
CRD+SPK	**α-NDCG@20**	**Loss**
SDM baseline	0.5528	-4.8%
DCS	0.6014	-3.3%
DCSF	0.6001	-3.9%
PM-2	0.6118	-3.9%
PM-2M	0.6137	-3.7%
CRD+SPK	**CPR@20**	**Loss**
SDM baseline	0.7818	-0.7%
DCS	**0.7968**	-1.3%
DCSF	0.8270	-0.8%
PM-2	0.8291	-0.6%
PM-2M	0.8294	-0.6%

Table 5.9. Results with collapsed dates for CRD+SPK with relative losses with respect to not collapsing dates. *All* entries *except* for the bold ones are significantly worse than their counterpart non-collapsed results (p-value < 0.02).

movie release dates in several countries and on different forms of media, or awards that were received for the movie. These are also relevant and thus yield a nice set of time intervals for diversification. The second example is topic number 867, about Cheney's hunting event. This happened on February 11, 2006, but many follow-up events can be counted: when the news was first reported by the ranch owner, when the incident report was issued, when Cheney talked about the event publicly for the first time etc. Topic number 867 with Cindy Sheehan is rather a borderline case between temporally ambiguous and unambiguous: she started her antiwar campaigns after her son's death in April 2004. The most noticed event for this topic is in August 2005, when Mrs Sheehan tried to meet the president at his residence. However her protests and demonstrations continued long afterward, which again is a nice source of time intervals for diversification. The next topic is a bit peculiar: brrreeeport. This

was an experiment by Robert Scoble from Microsoft in which bloggers were asked to add the word 'brrreeeport' to their posts to then see how quickly results would show up for that query in web search results. This happened in February 2006. We only have a handful of additional relevant times for this topic, and therefore this is the most unambiguous topic among the ones listed in Table 5.10. Another borderline case is topic number 1006, 'Mark Warner for President'. This happened during 2006, but the most outstanding related event happened in October 2006, when Mark Warner announced that he is not running for president. Finally, we have topic number 1015, Whole Foods wind energy, which mainly happened during 2006. This is rather a truly temporally unambiguous topic and it was hard for the annotators to find many relevant dates other than the main announcement and therefore the time intervals are rather broad.

Topic ID	Topic Title	Unambiguous Relevant Time	Some other related, relevant times
851	"March of the Penguins"	2005	2006; December 22, 2005; January 31, 2006; December 2005; January 2006; October 2005; August 2005; July 2005
867	cheney hunting	February 11, 2006	February 13, 2006; February 15, 2006; February 16, 2006; February 12, 2006; February 17, 2006
871	cindy sheehan	August 2005	2005; 2006; December 2005; February 1, 2006; November 2005
907	brrreeeport	February 2006	February 14, 2006; February 15, 2006; March 2006; 2006
1006	Mark Warner for President	October 2006	2005; 2004; November 2005; July 2005; January 2006; May 2005; 2006
1015	Whole Foods wind energy	2006	2005; January 2006; February 2006; September 2005; October 2005; November 2005; December 2005

Table 5.10. A few temporally unambiguous topics with their most outstanding "unambiguous" relevant time and some other related, relevant times for the topic.

We look at the straight-bias search results for these topics with the biases CRD + SPK and OTL+SLB, specifically observing the measures Precision-IA@20, α-NDCG@20, and CPR@20 in Table 5.11. For many of the queries we notice comparable or often much better performance than the reported average results in this chapter. The exceptions are: with CRD+SPK, Precision-IA@20 is low for 'Whole Foods wind energy', which is not surprising, given the vagueness of this topic with

Bias		CRD+SPK		
Topic ID	Topic Title	**Precision-IA@20**	**α-NDCG@20**	**CPR@20**
851	"March of the Penguins"	0.3552	0.6769	0.9763
867	cheney hunting	0.6005	0.6191	0.9189
871	cindy sheehan	0.4222	0.7177	0.9841
907	brrreeeport	0.3868	0.6373	0.8898
1006	Mark Warner for President	0.3565	**0.5136**	**0.6899**
1015	Whole Foods wind energy	**0.2921**	0.6646	0.8752
Average result		0.3267	0.6242	0.8335
Bias		OTL+SLB		
Topic ID	Topic Title	**Precision-IA@20**	**α-NDCG@20**	**CPR@20**
851	"March of the Penguins"	0.2701	0.4413	0.8513
867	cheney hunting	**0.0920**	**0.1512**	**0.4298**
871	cindy sheehan	0.2821	0.6162	0.8020
907	brrreeeport	0.3070	0.6264	0.8534
1006	Mark Warner for President	0.1822	0.4609	0.7552
1015	Whole Foods wind energy	0.3284	0.6528	0.8992
Average result		0.1562	0.4121	0.6769

Table 5.11. Some straight-bias results for queries from Table 5.10 with DCSF compared to average results over all queries.

respect to relevant times. 'Mark Warner for President' also poses a difficulty, and both CPR@20 and α-NDCG@20 are lower, but Precision-IA@20 is higher than the average. For the OTL+SLB bias combination, we have a surprising result: for the topic Cheney hunting we have lower performance for all three measures. Although we have many dates for this topic, emphasizing minority opinions with less important times seems to be a challenge. One possible explanation is that for this topic, the important dates are clustered mostly around February 2006. So, time-wise, the tail of the distribution for times is not very apart from the front. Hence, many documents mentioning tail times are very likely to also mention more popular times, which affects the overall bias calculations for the ranked list. This issue is rather specific to this topic, since the remaining results for OTL+SLB are all either comparable to the average or much better.

Overall, we did not see any evidence for temporally unambiguous queries performing noticeably worse than the average query. There are definitely some topics with fewer relevant times, and those may affect some of the measures, but we did not see any outstanding differences. This is good and may be attributed to the fact that we use document content dates spanning variable time intervals, which leaves more

flexibility with choosing very specific time points versus more vague, broader time intervals for a topic. This way, there is no reason to exclude temporally unambiguous queries for temporal diversification.

5.4.4.7 A Concrete Example

Lastly, we look at the topic 'women in Saudi Arabia.', number 1007, from the TREC Blog Track, diversified with the Crowd+Spike biases in Tables 5.12 and 5.13 using the DCSF algorithm. The aim here is to understand how diversifying for sentiment and time individually versus simultaneously can help. To fully focus on diversification without query aspect estimation errors, we use perfect data as in Section 5.4.4.3. In Table 5.12 we show two simultaneous combinations of sentiment+time: the upper example does this according to the provocativeness of the query's topic (see Section 5.4.3.5), which is in this case a 90% weight for sentiment, and 10% for time. The lower example shows an equal combination of the two dimensions. In Table 5.13 the results are shown for sentiment diversification and time diversification individually. We show brief excerpts or titles of the documents with the overall dominant sentiment in the "Sent." column, as well as a simplified representation of query time aspects in years only. "MO" in this column refers to "many other", meaning that this document contains many time mentions with a flat weight distribution.

The Crowd+Spike bias for this topic is 67% negative, 17% mixed/ neutral, and 16% positive for the sentiment dimension, and there are 21 relevant time intervals, some of which are shown in Table 5.14: the highest-weighted ones are 2006 with 34.4%, 2005 with 18.7%, January 2006 with 5.6%, and 2003 with 4.2% to name a few. The time intervals were manually judged as relevant, whereas the weights of the times are determined by their occurrence frequencies within all relevant documents for the query, as described in Section 5.4.3.3 ('Truth Judgments'). Among the 21 time intervals there are also less recent ones such as 1999 (1.6%), when women were

	Sentiment + Time Diversification, 90%-10%		
Rank	Excerpt	Sent.	Times
1	The Religious Policeman: Mutt the Muttawa	-	2002, 2005 - 2006
2	Happy Feminist: PROTESTING GENDER...	o	2000, 2002, 2005 - 2006
3	...when Saudi courts condemn women to death...	-	2003 - 2006
4	Orientalism and Islamophobia	o	1962, 1970, 1991, 2002 - 2004, 2006, MO
5	First women to win in Saudi elections	+	2004 - 2005
6	Laws discriminate against women...	-	1940, 1981, 1992, 1998, 2002 - 2006, MO
7	Saudi Arabia, Ever Our Friends And Allies	-	1981, 2001, 2005
8	Depressing Post: ... woman files a case against...	-	1960s, 1999, 2000 - 2005, MO
9	... scantily-clad women co-habitate...	-	1960, 1970, 1980s, 2005-2006
10	Their shabby treatment of women...	-	1978 - 1979, 2004, 2006

	Sentiment + Time Diversification, 50%-50%		
Rank	Excerpt	Sent.	Times
1	The Religious Policeman: Mutt the Muttawa	-	2002, 2005 - 2006
2	An option for Saudi Arabia	o	2003
3	...when Saudi courts condemn women to death...	-	2003 - 2006
4	Not a Desperate Housewife	-	1986, 2005
5	Soon, Saudi women may take the wheel	+	2005 - 2006
6	Saudi women continue to face serious obstacles...	-	2003 - 2005
7	Orientalism and Islamophobia	o	1962, 1970, 1991, 2002 - 2004, 2006, MO
8	Laws discriminate against women...	-	1940, 1981, 1992, 1998, 2002 - 2006, MO
9	Life in Saudi Arabia	o	2005 - 2006
10	Saudi Female Drive-a-thon Protest	-	1990, 2005

Table 5.12. CRD+SPK Bias with different Sentiment + Time combinations: Top 10 results for DCSF model for query number 1007, 'women in Saudi Arabia.' - denotes a negative document, o refers to mixed/neutral, and + to positive.

	Sentiment Diversification		
Rank	Excerpt	Sent.	Times
1	The Religious Policeman: Mutt the Muttawa	-	2002, 2005 - 2006
2	Happy Feminist: PROTESTING GENDER...	o	2000, 2002, 2005 - 2006
3	Saudi Arabia, Ever Our Friends And Allies	-	1981, 2001, 2005
4	Orientalism and Islamophobia	o	1962, 1970, 1991, 2002 - 2004, 2006, MO
5	First women to win in Saudi elections	+	2004 - 2005
6	Laws discriminate against women...	-	1940, 1981, 1992, 1998, 2002 - 2006, MO
7	Depressing Post: ...woman filed a case against...	-	1960s, 1999, 2000 - 2005, MO
8	Their shabby treatment of women...	-	1978 - 1979, 2004, 2006
9	Oprah is being smuggled into Saudi Arabia...	-	2002 - 2005
10	Between tradition and demands for change	o	1979, 1980s, 1990s, 2003 - 2005

	Time Diversification		
	Excerpt	Sent.	Times
1	The Religious Policeman: Mutt the Muttawa	-	2002, 2005 - 2006
2	An option for Saudi Arabia	o	2003
3	Not only did women vote in the elections...	+	2005
4	Soon, Saudi women may take the wheel	+	2005 - 2006
5	...when Saudi courts condemn women to death...	-	2003 - 2006
6	First women to win in Saudi elections	+	2004 - 2005
7	The War to Mobilize Democracy	+	1981, 2003, 2005 - 2006
8	...rights of women to participate in elections...	+	2004 - 2005, 2009
9	Life in Saudi Arabia	o	2005 - 2006
10	Being a Child in Saudi Arabia	o	2003 - 2006

Table 5.13. CRD+SPK Bias with Sentiment Diversification and Time Diversification individually: Top 10 results for DCSF model for query number 1007, 'women in Saudi Arabia.' - denotes a negative document, o refers to mixed/neutral, and + to positive.

Time	Weight
2006	34.4%
2005	18.7%
January 2006	5.6%
2003	4.2%
2004	3.8%
December 2005	3.0%
2002	3.0%
2000	3.0%
...	...
1999	1.6%
2008	1.1%
1970s	0.6%
2009	0.2%

Table 5.14. Some truth times and weights for query number 1007, 'women in Saudi Arabia.'

allowed to attend the Saudi council for the first time, and the 1970s (0.6%), which is when many women's institutions such as universities and colleges were established. Clearly, more 'recent times' given the corpus spanning 2003-2006 are more important for this topic when events happened related to women's rights to drive, vote, work with men etc. Looking at Table 5.13, we can see in the Time Diversification results that these more recent times are preferred. These have larger weights and thus satisfy the SPK bias well: documents heavily mentioning these times only are preferred than others. Note that the overall *sentiment* bias in that list is bad: we have 5 positive documents (50%), 2 negative ones (20%), and 3 mixed/neutral ones (30%), which is very far from our desired CRD bias 67%-16%-17% for sentiments. The other three results have much better sentiment biases: Sentiment diversification and the 50%-50% sentiment + time combination both achieve 60%-10%-30%, whereas the 90%-10% combined sentiment + time diversification result achieves 70%-10%-20%, which is the closest to the CRD bias. These three results also achieve a better variety with query time aspects, including documents referring to the 1970s, 1986, and 1999, while maintaining an overall emphasis on the more recent times as requested by the

SPK bias. We conclude that because the Query Time Aspects Distribution is much flatter and there are many (21!) aspects, this dimension is much easier to address, whereas for the sentiment dimension – given its skewness – the desired bias is harder to achieve if it is not explicitly considered. Hence, Time Diversification on its own cannot replace a biased sentiment + time diversification. The advantage of using the latter over individual sentiment diversification is that with the linear combination we can control the influence of each result while maintaining the overall desired bias. This way, the results for the 50%-50% combination look similar to the individual Time and Sentiment Diversification results without much sacrificing the overall bias in the list. The 90%-10% combined result is by nature similar to Sentiment Diversification individually, but the former includes 1 result at rank 3 from Time Diversification (individually) which is not present in the 'sentiment only' results. This is a desired effect for this 90%-10% combination. We can see that the user has control over the results by not only specifying the desired biases, but also over the extent to which a certain dimension shall be emphasized. This is achieved without the user knowing any details about the topic such as pre-existing sentiment biases, crucial times etc. – which are all handled by the system.

5.5 Summary

In this chapter we first motivate the need to study multiple non-topical dimensions together for better understanding position on a controversial topic. To back this up, we perform analyses on two publicly available query logs, hinting at evidence about peoples' interest in subjective and temporal information about controversial topics. On the diversification front, we present a general bias framework to be used during diversification with non-topical aspects of several dimensions. This framework is used together with different diversification algorithms for the dimensions sentiment and time. In the diversification experiments, we choose three extreme target biases

along the continuum for each dimension to explore their impact. We discover the following:

1. For all different bias combinations for sentiments and time, the DCSF, PM-2, and PM-2M models perform best and prove most stable with noisy sentiment and time labels. While α-NDCG is decent for the DCS model, it only performs consistently well if the labels are perfect.

2. If equal diversification is used for both dimensions in case a dataset for estimating biases is not available, we observe significant losses over four different bias combinations.

3. We compare the results from (1) to results with perfect labels for the three base bias cases to see the maximum performance under ideal circumstances.

4. We view query time aspects at different granularities by collapsing overlapping times and their weights. This particularly helps the Slab bias.

5. We also observe some temporally unambiguous queries individually and compare their results to the reported average. We do not find any indication for noticeable different results for this class of queries.

6. Finally, by means of an example we demonstrate how diversifying simultaneously for several dimensions can be more helpful than diversifying individually for each.

Apart from investigating simultaneously diversifying for several dimensions, another contribution in this work is temporal diversification with time expressions extracted from within documents.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

In this thesis, we investigated how to integrate non-topical aspects into information retrieval. For this, we chose an information retrieval task, search result diversification, and non-topical aspects from two different dimensions: opinionatedness and time. First, we considered the characteristics of each dimension and its query aspects in isolation of the applied information retrieval task: what is opinionatedness? What is time? To support this discussion, we devised measures at the topic or document level to quantify the non-topical nature of a unit of text, and applied them to the TREC Blog Track dataset. For opinionatedness, these measures were provocativeness, balance, and average sentiment of a topic. Additionally, the relationship between sentiments and a document can be expressed by means of a sentiment score. For time, we devised a measure that considers the importance of a time (interval) to a document and how well relevant time intervals in the document cover the time interval in question. We combined the sentiment and time measures into an interestingness measure. The purpose of these measures was to show how retrieved documents for a query can be reranked in order of relevance to sentiments, time, or a combination of both dimensions.

Then, we focused on the chosen information retrieval task for diversifying search results with non-topical aspects of one or several dimensions. For this, we extended two existing diversification frameworks, xQuAD (R. L. Santos et al., 2010a) and Diversity by Proportionality (Dang & Croft, 2012), to work with the sentiment and

time dimensions. We used fixed query sentiment aspects, positive, negative, and neutral, and a variable number of time aspects for each query, extracted from within documents and from Wikipedia for diversification. We also proposed variations to these algorithms: xQuAD was modified in Sections 3.2.2.2.2 and 4.4.1.1.2 to not only use the strength of sentiment (or relevance) scores as originally defined, but to also consider the frequency of such query sentiment aspects to control diversity. For the proportionality models, we proposed a variation to PM-2 (Dang & Croft, 2012) that adapts the quotient calculation in case there are not enough documents present in the retrieved list for a certain query aspect. Our experiments with the sentiment dimension only on the TREC Blog Track revealed that the proportionality based models and SCSF significantly outperform the SDM baseline and the xQuAD-like approach SCS for most measures and sufficiently high sentiment classification accuracies. In the experiments with the time dimension only, we noted that the Wikipedia aspect weights proved rather noisy, with most methods only marginally improving the results over the SDM baseline, and sometimes over TCS, the time-adapted xQuAD-like SCS model. The oracle results with perfect labels were much better, showing the maximum potential for improvement for our models. Surprisingly, TCS sometimes achieved the best results, so we hypothesized that this model may be more suitable for dimensions with many aspects and rather flatly distributed query aspect weights.

In the "interestingness" experiments with both dimensions, sentiments and time, we observed several outcomes:

1. with the introduction of the second dimension, time, diversification is more sensitive to the sentiment classifier's accuracy: for some measures, only higher accuracies yielded significant diversification performance when compared to the SDM baseline or the DCS model;

2. in general, again the proportionality based models and DCSF (generalized SCSF) performed best, however with perfect labels DCS (generalized SCS) also proved to be effective;

3. there does not seem to be a noticeable difference between the performance for temporally ambiguous versus unambiguous queries, which we demonstrated by means of example queries.

We also introduced a general bias framework to be used for non-topical diversification. This seamlessly integrates the three target biases that we first defined for sentiments to work with any dimension with a fixed or variable and finite number of query aspects. The bias framework is applied during diversification and evaluation to indicate which bias shall be considered for rearranging the results. For each dimension a different bias can be chosen. We presented two variations to estimating the Outlier or Slab biases: one inverts the original query dimension aspects distribution, the other reverts it. Our experiments evaluated the efficiency of the diversification frameworks, algorithms, and biases given different kinds of settings: noisy query aspect labels, and perfect query aspect labels. We also simulated the lack of data and the effect of substituting biases for one another. In the experiments on the TREC Blog Track, we diversified across a single dimension as well as across multiple dimensions. Overall, we made the following additional observations:

1. whether we diversify with one dimension or several, using the intended bias is crucial: we observed significant losses for the DCSF and PM-2 models when the intended bias was substituted with equal diversification – on average the loss of performance was over 10%;

2. we viewed query time aspects at different granularities by collapsing overlapping times and their weights. This particularly helped the Slab bias, boosting weights

for smaller time intervals, which are rather present in the tail of Query Time Aspects Distributions, which are emphasized through the Slab bias.

For the evaluations, we adapted existing intent-aware evaluation measures to work with multiple dimensions and biases.

6.2 Future Work

There are many interesting directions for future work. First, the ideas presented in this work are not only valuable for sentiment or time diversity, but they can also be applied to topical diversity with modifications. To what extent does it make sense to consider biases for topical diversity? For instance, with an Outlier bias-like approach, underrepresented query sentiment aspects could be highlighted in search results. Further, we have proposed different extensions to existing diversification models such as xQuAD and PM-2 with the SCSF/DCSF and PM-2M models, which may be effective for topical diversity as well.

For opinionatedness in particular, during diversification we mainly focused on query sentiment aspects. It would be interesting to analyze opinion or topical arguments and sentiments together with biases for diversification. One question to solve is what kind of biases could be defined to capture both, and whether more fine-grained topic-specific biases would be required. For this type of 'joint modeling' of topical and non-topical dimensions for diversification, there are in general two approaches: either one subtopic can first be chosen or 'fixed', within which then diversification is achieved across different sentiments; or, search results can be diversified across different subtopics and sentiments simultaneously. Unfortunately, this requires special datasets with judgments supporting such research, which is why we have not been able to pursue it. In this thesis, we mainly used the TREC Blog Track dataset as a basis for experiments following prior work (Demartini & Siersdorfer, 2010; Demartini, 2011). Sentiment classifier training was also done on documents retrieved by a com-

mercial search engine. Alternatively, we have also considered the ClueWeb09 corpus[1], but due to the lack of judgments or labels, these corpora did not prove useful for our diversification experiments or sentiment classifier training. Therefore, the creation of suitable datasets or the addition of labels to existing datasets would greatly benefit single or multiple dimension diversification research.

There is a lot of scope for future work in Temporal or Time Diversification: for example, exploring different kinds of biases that explicitly emphasize recency or freshness versus background information. In Chapter 5 we tried one alternative way of working with times and dates by collapsing them, but others could be explored in addition. Perhaps, classifiers can be trained for categorizing detected times and dates for more fine-grained diversification. For example, it may be useful to know whether a time mention in a document comes from a communication, or whether it is a publication date mentioned in the document, or whether it is mentioned as a fact such as in news. Since frequency-based weights for Wikipedia time aspects proved rather noisy, alternatives for estimating weights better can be explored as well.

Other dimensions with non-topical aspects can be experimented with than sentiments and time. Often, we have mentioned geography as one such suitable candidate. One major issue in this regard is the clarification of how and where to obtain relevance judgments and data for new dimensions, as mentioned above. Our diversification frameworks are designed with two different kinds of dimensions in mind: those with a variable but finite number of aspects across queries or topics such as time, and those with a fixed number of aspects such as sentiments. Other (non-topical) dimensions should fall into one of these categories. Dimensions with an infinite number of aspects would require a reasonable upper bound for equal diversification or for the reversal of values in the query aspects distribution. In Chapter 4 we mentioned the

[1]http://lemurproject.org/clueweb09/

potential of the TCS method to perform well for dimensions with many aspects, so this is something to explore.

Another interesting question is the personalization aspect for diversifying search results across multiple dimensions and biases: can we, given a user and her query, predict, which dimensions and which bias combinations will be most useful to her? Of course, this requires implementing the system as part of a search engine or other interface, with which then some user data can be collected. This would then serve as the underlying source of information for investigating this question.

We discovered in Chapter 3 that 3-class sentiment classification for document-length blogs is a difficult task, and therefore, we had to simulate a sentiment classifier for the experiments. This is a natural language processing task that definitely requires attention. Most sentiment classification research has ubeen done on short texts spanning a single sentence or paragraph like tweets or movie reviews (Pang et al., 2002; Turney, 2002). This is a more straightforward task than full-document sentiment classification since smaller units of text contain fewer sentiments than full-length documents. Still, there are substantial challenges to be resolved in this area, like the detection of sentiments for irony and sarcasm (Davidov et al., 2010; González-Ibáñez et al., 2011). Can the accuracy of current state-of-the-art sentiment classification tools on longer documents be improved? If this is possible, applying our techniques at web scale becomes much more realistic. In case this is difficult to realize, the question is reduced to how the diversification models can be adapted to handle noisy classification input. Again, one of the challenges in improving document-length sentiment classification is the lack of datasets on which classifiers can be trained and tested. The TREC Blog Track is the only major available dataset with opinion and relevance judgments consisting of blogs, used by prior work for sentiment classification training and testing (Demartini & Siersdorfer, 2010; Demartini, 2011). Other work has used web data (Kacimi & Gamper, 2011) or the ClueWeb09 dataset (Vural,

Cambazoglu, & Senkul, 2012), but the performance of their classifiers is not specified. So research in this area would greatly benefit from the creation of additional datasets and/or the markup of existing datasets with sentiment labels.

REFERENCES

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

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). New York, NY, USA: ACM.

Aktolga, E., & Allan, J. (2014). Diversifying Across Non-Topical Aspects with Biases for Several Dimensions. In *submission.*

Allan, J. (Ed.). (2002). *Topic Detection and Tracking: event-based information organization*. Norwell, MA, USA: Kluwer Academic Publishers.

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). New York, NY, USA: ACM.

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).

Alonso, O., & Gertz, M. (2006). Clustering of search results using temporal attributes. In *Proceedings of the 29th annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 597–598). New York, NY, USA: ACM.

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). New York, NY, USA: ACM.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In N. C. C. Chair) et al. (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA).

Berberich, K., & Bedathur, S. (2013). Temporal Diversification of Search Results. In *SIGIR 2013 Workshop on Time-aware Information Access (TAIA).*

Berberich, K., Bedathur, S., Alonso, O., & Weikum, G. (2010). A language modeling approach for temporal information needs. In *Proceedings of the 32nd European conference on Advances in Information Retrieval* (pp. 13–25). Berlin, Heidelberg: Springer-Verlag.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the seventh international conference on World Wide Web 7* (pp. 107–117). Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V.

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). New York, NY, USA: ACM.

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). New York, NY, USA: ACM.

Cartright, M.-A., Aktolga, E., & Dalton, J. (2009). Characterizing the subjectivity of topics. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 642–643). New York, NY, USA: ACM.

Chasin, R. (2010). Event and Temporal Information Extraction towards Timelines of Wikipedia Articles. *Simile*, 1–9.

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). New York, NY, USA: ACM.

Chen, S. F., & Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics* (pp. 310–318). Stroudsburg, PA, USA: Association for Computational Linguistics.

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). New York, NY, USA: ACM.

Clarke, C. L., Kolla, M., & Vechtomova, O. (2009). An Effectiveness Measure for Ambiguous and Underspecified Queries. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory* (pp. 188–199). Berlin, Heidelberg: Springer-Verlag.

Clarke, C. L. A., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 Web Track. In *Proc. of TREC-2009.* Retrieved from `http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf`

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory.* New York, NY [u.a.]: Wiley.

Dai, N., & Davison, B. D. (2010). Capturing Page Freshness for Web Search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 871–872). New York, NY, USA: ACM.

Dai, N., Shokouhi, M., & Davison, B. D. (2011). Learning to rank for freshness and

relevance. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 95–104). New York, NY, USA: ACM.

Dang, V., & Croft, W. B. (2012). Diversity by proportionality: an election-based approach to search result diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 65–74). New York, NY, USA: ACM.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning* (pp. 107–116). Stroudsburg, PA, USA: Association for Computational Linguistics.

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). Berlin, Heidelberg: Springer-Verlag.

Demartini, G., & Siersdorfer, S. (2010). Dear search engine: what's your opinion about...?: sentiment analysis for semantic enrichment of web search results. In *Proceedings of the 3rd International Semantic Search Workshop* (pp. 4:1–4:7). New York, NY, USA: ACM.

Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., ... Zha, H. (2010). Time is of the essence: improving recency ranking using Twitter data. In *Proceedings of the 19th international conference on World wide web* (pp. 331–340). New York, NY, USA: ACM.

Eguchi, K., & Lavrenko, V. (2006). Sentiment retrieval using generative models. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 345–354). Stroudsburg, PA, USA: Association for Computational Linguistics.

Elsas, J. L., & Dumais, S. T. (2010). Leveraging Temporal Dynamics of Document Content in Relevance Ranking. In *Proceedings of the third acm International conference on Web search and data mining* (pp. 1–10). New York, NY, USA: ACM.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, *9*, 1871–1874.

Feige, U. (1998, July). A Threshold of Ln N for Approximating Set Cover. *J. ACM*, *45*(4), 634–652.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 363–370). Stroudsburg, PA, USA: Association for Computational Linguistics.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987, November). The vocabulary problem in human-system communication. *Commun. ACM*, *30*(11), 964–971.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, *38*.

Gerani, S., Carman, M. J., & Crestani, F. (2009). Investigating Learning Approaches for Blog Post Opinion Retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval* (pp. 313–324). Berlin, Heidelberg: Springer-Verlag.

Girill, T. R. (1985). Online access AIDS for documentation: a bibliographic outline. *SIGIR Forum*, *18*, 24–27.

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying Sarcasm in Twitter: A Closer Look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2* (pp. 581–586). Stroudsburg, PA, USA: Association for Computational Linguistics.

Harman, D. (2002). Overview of the TREC 2002 Novelty Track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002), NIST Special Publication 500-251* (pp. 46–55).

He, B., Macdonald, C., & Ounis, I. (2008). Ranking opinionated blog posts using OpinionFinder. In *Proceedings of the 31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 727–728). New York, NY, USA: ACM.

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). New York, NY, USA: ACM.

Hersh, W., & Over, P. (1999). TREC-8 interactive track. *SIGIR Forum*, *33*, 8–11.

Hersh, W., Turpin, A., Price, S., Chan, B., Kramer, D., Sacherek, L., & Olson, D. (2000). Do batch and user evaluations give the same results? In *Proceedings of the 23rd annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 17–24). New York, NY, USA: ACM.

Hilderman, R. J., & Hamilton, H. J. (2003). Measuring the interestingness of discovered knowledge: A principled approach. *Intell. Data Anal.*, *7*, 347–382.

Huang, X., & Croft, W. B. (2009). A unified relevance model for opinion retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 947–956). New York, NY, USA: ACM.

Hurst, M., & Nigam, K. (2004). Retrieving topical sentiments from online document collections. In *In Document Recognition and Retrieval XI* (pp. 27–34).

Jatowt, A., Kawai, Y., & Tanaka, K. (2005). Temporal Ranking of Search Engine Results. In A. H. H. Ngu, M. Kitsuregawa, E. J. Neuhold, J.-Y. Chung, & Q. Z. Sheng (Eds.), *WISE* (Vol. 3806, p. 43-52). Springer.

Jia, L., Yu, C., & Meng, W. (2009). The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1827–1830). New York, NY, USA: ACM.

Jones, R., & Diaz, F. (2007). Temporal profiles of queries. *ACM Trans. Inf. Syst.*, *25*(3).

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). New York, NY, USA: ACM.

Kacimi, M., & Gamper, J. (2012). MOUNA: mining opinions to unveil neglected arguments. In *Proceedings of the 21st acm International conference on Information and knowledge management* (pp. 2722–2724). New York, NY, USA: ACM.

Kanhabua, N., & Nørvåg, K. (2010). Determining time of queries for re-ranking search results. In *Proceedings of the 14th European conference on Research and advanced technology for digital libraries* (pp. 261–272). Berlin, Heidelberg: Springer-Verlag.

Kanhabua, N., Romano, S., & Stewart, A. (2012). Identifying Relevant Temporal Expressions for Real-World Events. In *SIGIR 2012 Workshop on Time-aware Information Access (TAIA).*

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). New York, NY, USA: ACM.

Keikha, M., Gerani, S., & Crestani, F. (2011a). TEMPER: a Temporal Relevance Feedback Method. In *Proceedings of the 33rd European conference on Advances in information retrieval* (pp. 436–447). Berlin, Heidelberg: Springer-Verlag.

Keikha, M., Gerani, S., & Crestani, F. (2011b). Time-based Relevance Models. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1087–1088). New York, NY, USA: ACM.

Kulkarni, A., Teevan, J., Svore, K. M., & Dumais, S. T. (2011). Understanding Temporal Query Dynamics. In *Proceedings of the fourth acm International conference on Web search and data mining* (pp. 167–176). New York, NY, USA: ACM.

Kuzey, E., & Weikum, G. (2012). Extraction of Temporal Facts and Events from Wikipedia. In *Proceedings of the 2nd Temporal Web Analytics Workshop* (pp. 25–32). New York, NY, USA: ACM.

Lee, Y., Na, S.-H., Kim, J., Nam, S.-H., Jung, H.-Y., & Lee, J.-H. (2008). KLE at Trec 2008 Blog Track: Blog post and feed retrieval. In *Proceedings of TREC-08.*

Li, X., & Croft, W. B. (2003). Time-based Language Models. In *Proceedings of the twelfth International conference on Information and knowledge management* (pp. 469–475). New York, NY, USA: ACM.

Macdonald, C., Ounis, I., & Soboroff, I. (2007). Overview of the TREC-2007 Blog Track. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings.*

Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th International conference on World Wide Web* (pp. 171–180). New York, NY, USA: ACM.

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). New York, NY, USA: ACM.

Metzler, D., Jones, R., Peng, F., & Zhang, R. (2009). Improving Search Relevance

for Implicitly Temporal Queries. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 700–701). New York, NY, USA: ACM.

Nam, S.-H., Na, S.-H., Lee, Y., & Lee, J.-H. (2009). DiffPost: Filtering Non-relevant Content Based on Content Difference between Two Consecutive Blog Posts. In M. Boughanem, C. Berrut, J. Moth, & C. Soul-Dupuy (Eds.), *ECIR* (Vol. 5478, p. 791-795). Springer.

Nunes, S., Ribeiro, C., & David, G. (2008). Use of Temporal Expressions in Web Search. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, & R. White (Eds.), *Advances in Information Retrieval* (Vol. 4956, p. 580-584). Springer Berlin Heidelberg.

Ounis, de Rijke, M., Macdonald, C., Mishne, G., & Soboroff. (2006). Overview of the TREC-2006 Blog Track. In *Proceeddings of the 15th Text REtrieval Conference (TREC 2006).*

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10* (pp. 79–86). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pon, R. K., Cardenas, A. F., Buttler, D., & Critchlow, T. (2007). Tracking multiple topics for finding interesting articles. In *Proceedings of the 13th ACM SIGKDD International conference on Knowledge discovery and data mining* (pp. 560–569). New York, NY, USA: ACM.

Pon, R. K., Cárdenas, A. F., Buttler, D. J., & Critchlow, T. J. (2011). Measuring the interestingness of articles in a limited user environment. *Inf. Process. Manage.*, *47*, 97–116.

Radlinski, F., Szummer, M., & Craswell, N. (2010). Metrics for assessing sets of subtopics. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 853–854). New York, NY, USA: ACM.

Robertson, S. E. (1977). The Probability Ranking Principle in IR. *Journal of Documentation*, *33*(4), 294–304.

Sakai, T., & Joho, H. (2011). *Overview of NTCIR-9.*

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). New York, NY, USA: ACM.

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). New York, NY, USA: ACM.

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). New York, NY, USA: ACM.

Santos, R. L. T., Macdonald, C., McCreadie, R., Ounis, I., & Soboroff, I. (2012). Information Retrieval on the Blogosphere. *Found. Trends Inf. Retr.*, *6*(1), 1–125.

Sato, N., Uehara, M., & Sakai, Y. (2003). Temporal ranking for fresh information retrieval. In *Proceedings of the sixth International workshop on Information retrieval with Asian languages - Volume 11* (pp. 116–123). Stroudsburg, PA, USA: Association for Computational Linguistics.

Sato, N., Uehara, M., & Sakai, Y. (2004). FTF IDF Scoring for Fresh Information Retrieval. In *Proceedings of the 18th International Conference on Advanced Information Networking and Applications - Volume 2* (pp. 165–). Washington, DC, USA: IEEE Computer Society.

Seki, K., & Uehara, K. (2009). Adaptive subjective triggers for opinionated document retrieval. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 25–33). New York, NY, USA: ACM.

Soboroff, I., & Harman, D. (2005). Novelty detection: the TREC experience. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 105–112). Stroudsburg, PA, USA: Association for Computational Linguistics.

Swan, R., & Allan, J. (1998). Aspect windows, 3-d visualizations, and indirect comparisons of information retrieval systems. In *Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 173–181). New York, NY, USA: ACM.

Swan, R., & Allan, J. (2000). Automatic Generation of Overview timelines. In *Proceedings of the 23rd annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 49–56). New York, NY, USA: ACM.

Tan, P.-N., Kumar, V., & Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD International conference on Knowledge discovery and data mining* (pp. 32–41). New York, NY, USA: ACM.

Tsytsarau, M., Palpanas, T., & Denecke, K. (2010). Scalable discovery of contradictions on the web. In *Proceedings of the 19th International conference on World wide web* (pp. 1195–1196). New York, NY, USA: ACM.

Turney, P. D. (2002). Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 417–424). Stroudsburg, PA, USA: Association for Computational Linguistics.

Uehara, M., & Sato, N. (2005). Information Retrieval based on Temporal Attributes in WWW Archives. In *Proceedings of the 11th International Conference on Parallel and Distributed Systems - Volume 01* (pp. 756–761). Washington, DC, USA: IEEE Computer Society.

Vallet, D., & Castells, P. (2012). Personalized diversification of search results. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 841–850). New York, NY, USA: ACM.

Verberne, S. (2011). In search of the Why: developing a system for answering why-questions. *SIGIR Forum*, *44*, 90–90.

Vural, A. G., Cambazoglu, B. B., & Senkul, P. (2012). Sentiment-focused Web Crawl-

ing. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (pp. 2020–2024). New York, NY, USA: ACM.

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). New York, NY, USA: ACM.

Xu, X., Meng, T., Cheng, X., & Liu, Y. (2011). A probabilistic model for opinionated blog feed retrieval. In *Proceedings of the 20th International conference companion on World wide web* (pp. 155–156). New York, NY, USA: ACM.

Yu, P. S., Li, X., & Liu, B. (2004). On the Temporal Dimension of Search. In *Proceedings of the 13th International World Wide Web conference on Alternate track papers & posters* (pp. 448–449). New York, NY, USA: ACM.

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

Zhang, W., Jia, L., Yu, C., & Meng, W. (2008). Improve the effectiveness of the opinion retrieval and opinion polarity classification. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 1415–1416). New York, NY, USA: ACM.

Zhang, W., Yu, C., & Meng, W. (2007). Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 831–840). New York, NY, USA: ACM.