NOVELTY AND DIVERSITY IN SEARCH RESULTS

by

Praveen Chandar

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer and Information Sciences

Summer 2014

NOVELTY AND DIVERSITY IN SEARCH RESULTS

by

Praveen Chandar

Approved: _____

Errol Lloyd, Ph.D.
Chair of the Department of Computer and Information Sciences

Approved: _____

Babatunde Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____

James G. Richards, Ph.D.
Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Benjamin A. Carterette, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Hui Fang, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Vijay K. Shanker, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Kathleen F. McCoy, Ph.D.
Member of dissertation committee

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Benjamin Carterette, for his support and guidance through the course of my grad school experience. His suggestions and constant words of encouragement were absolutely essential in making this work possible. I had the freedom to work independently and could rely on his guidance whenever needed. I consider myself lucky to have worked with him; he has always been patient in answering my questions, and extremely helpful. His guidance and advice helped shape my path towards a promising career, and for that I thank him sincerely.

I would like to thank my committee members: Vijay Shanker, Kathleen McCoy, and Hui Fang. They all had a unique perspective and provided valuable feedback, which was crucial for this project. I thank them for the careful reviews and suggestions provided during various stages of the work. Also, I have been fortunate to have worked with William Webber and Hema Raghavan, who helped me improve as a researcher.

I am also grateful to all past and present members of my research group for providing their feedback and input to improve this work. I would especially like to thank Keith Trnka, Rich Burns, Oana Tudor, Charlie Greenbacker, Peng Wu, Dan Blanchard and Rashida Davis for their help during my early days of grad school. Towards the latter part of my Ph.D., I was fortunate to share the lab space with Priscilla Moraes, Dongqing Zhu, Ashwani Rao, Ivanka Li, Ashraf Bah and Mustafa Zengin. I definitely learned a lot from our discussions over lunch and coffee breaks; I am glad to have interacted with you all.

To all my friends in Delaware and outside, my grad school life would not have been the same without you all. My friends helped me stay focused and were supportive during the tough times. I have enjoyed their company and greatly value their friendship.

I would also like to extend my deepest gratitude to my family, without whom this dissertation would not have been possible. My father has been a huge inspiration throughout my life; his encouragement and motivation played a large role in my decision to pursue higher studies. I thank my mother for her patience and endless support throughout my life, for which I will be forever indebted. I would like to dedicate this dissertation to my parents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Information retrieval (IR) is the process of obtaining relevant information for a given information need. The concept of relevance and its relation to information needs is of central concern to IR researchers. Until recently, much work in IR settled with a notion of relevance that is *topical* — that is, containing information "about" a specified topic — and in which the relevance of a document in a ranking is independent of the relevance of other documents in the ranking. But such an approach is more likely to produce a ranking with a high degree of redundancy; the amount of novel information available to the user may be minimal as they traverse down a ranked list.

In this work, we focus on the *novelty and diversity* problem that models relevance of a document taking into account the inter-document effects in a ranked list and diverse information needs for a given query. Existing approaches to this problem mostly rely on identifying *subtopics* (disambiguation, facets, or other component parts) of an information need, then estimating a document's relevance independently w.r.t each subtopic. Users are treated as being satisfied by a ranking of documents that covers the space of subtopics as well as covering each individual subtopic sufficiently.

We propose a novel approach that models novelty implicitly while retaining the ability to capture other important factors affecting user satisfaction. We formulate a set of hypotheses based on the existing subtopic approach and test them with actual users using a simple conditional preference design: users express a preference for document A or document B *given* document C. Following this, we introduce a novel triplet framework for collecting such preference judgments and using them to estimate the total utility of a document while taking inter-document effects into account. Finally, a set of utility-based metrics are proposed and validated to measure the effectiveness of a system for the novelty and diversity task.

Chapter 1

INTRODUCTION

As humans, we spend much of our time exploring and seeking information to learn more about our environment. Information seeking starts with a need for information that is often intangible, making it hard to specify in natural language [16]. The information need is a representation of a problem as perceived in a user's mind and is the starting point of all information searches. It is usually translated into natural language or a set of keywords, which is input to a retrieval system that searches a collection of documents for the required information. An information retrieval system often returns a ranked list of documents that are subject to self-evaluation by the user. A decision of whether to stop or to continue the search is made upon examining the information present in each document. The goal of an information retrieval (IR) system is to efficiently and effectively provide information relevant to the user's needs, before they abandon the search.

The decision to abandon or continue with the search process depends on many factors, including the relevance of the documents examined. Each document examined also affects the existing knowledge of the user, thereby potentially altering the information need in subtle, hard-to-quantify ways. Therefore, the concept of relevance and its relation to information need is of central concern to building and evaluating IR systems [159]. The definition of relevance has taken various forms; and various kinds of relevance from different viewpoints have been proposed in the past [107]. Among these definitions, the need for *novelty and diversity* in a ranked list along with relevance was addressed by Goffman [73]. He stressed that the relevance of a document in a ranked list, along with topical relevance, is dependent on the previous documents retrieved.

Until recently, most researchers settled with the notion of relevance in which the relevance of a document was independent of other documents in the ranking. The *Probability Ranking Principle* (PRP) proposed by Robertson [130], a key principle guiding the development of retrieval systems for nearly forty years, is based on independent relevance. The PRP says optimal performance is achieved when documents are ranked in decreasing order of probability of relevance, under the assumption that documents are independently relevant. A major problem with this approach is that it encourages redundancy and thereby potentially reduces the amount of novel information available to the user.

Some information needs could be decomposed into several distinct, smaller pieces of information sometimes called *subtopics* [58, 53, 117]. Often, simple tasks like finding a home page require only a single *subtopic* — one that represents the page to be found — to satisfy the underlying information need, whereas more complex tasks such as writing a report or planning a trip may require searching for many subtopics: in the case of planning a vacation, subtopics may include possible destinations, things to do in each of those destinations, places to stay, travel packages, and more. This representation is more suitable to describe our novelty and diversity task: an information need can be represented by a set of subtopics that are clearly delineated such that there may be many documents relevant to a single subtopic, and a single document may be relevant to many subtopics.

In a typical retrieval scenario, users interact with the IR systems by formulating their information needs in the form of a query (one or more keywords). The query acts as an input to IR systems that produces a ranking of documents as an output, the idea being that documents ranked higher are more likely to satisfy the user's information need. Note that the information need is expressed in the form of a query, and two or more information needs can be represented by the same query (*i.e.* the query might be ambiguous) or users might be interested in different subtopics of the information need.

In this work, we propose ways to solve the above mentioned problem by redefining the notion of relevance and defining new evaluation measures. We explain the

various components required to study the novelty and diversity problem as follows:

1. *Understanding User Requirements:* Identifying and analyzing factors that influence user preferences is key to building and evaluating retrieval systems. Steps must be taken to answer questions such as *why does a user prefer one document to the other?* and *what factors play a role in such preferences?*.

2. *Modeling:* The novelty of a document depends on documents that have been ranked above it, and cannot be based solely on probability of relevance. Thus, identifying factors affecting novelty is key to building models for this problem.

3. *Evaluation:* Judgments based solely on binary or graded judgments are made considering individual documents and are not dependent on other documents. Therefore, new evaluation measures need to be developed that account for various novelty factors. This is the primary focus of this work.

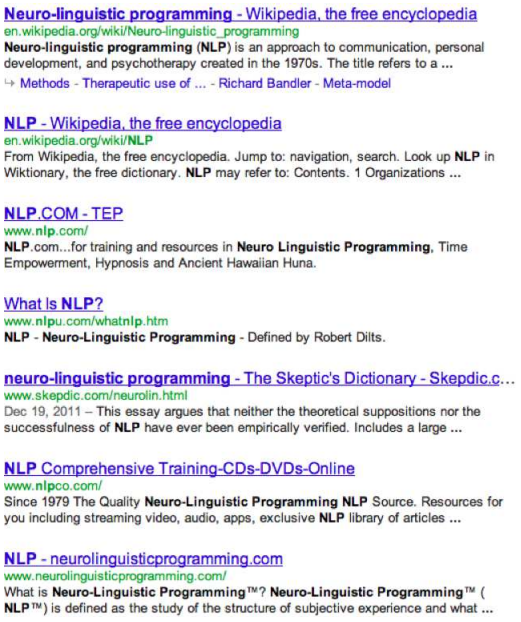

Figure 1.1: Search Results for the query 'nlp' on Google. Documents in the top results cover a single subtopic.

1.1 Novelty and Diversity

The novelty and diversity tasks requires systems to deal with the diverse information needs of the users while reducing redundancy in the ranked list. In order to

understand the task better, let us consider a few search scenarios that emphasize the need for novel and diverse information in the ranked list.

Scenario 1

First, consider a user looking for information on *natural language processing* who inputs the query *nlp*. Figure 1.1 shows the top results for the query *nlp* by a popular commercial web search engine. It can be seen that all documents contain information about *neuro-linguistic programming*. The ranking may be desirable for a user interested in *neuro-linguistic programming*, but it is not at all useful for a user looking for information on *natural language processing*. Since the query *nlp* is ambiguous and could potentially have several information needs associated with it, systems could improve overall user satisfaction by covering both information needs in their ranking; a ranking in which the first document covers the topic *neuro-linguistic programming* and the second covers the topic *natural language processing* would likely be superior to a ranking in which both top documents are about either one of those topics.

Scenario 2

Next, consider a scenario in which a user has an unambiguous but broad information need such as *how to build a wooden fence*. The user would likely look for various pieces of information such as *ways to build a fence*, *materials for building fences*, *where to buy materials for building a fence*, etc. Assuming the user traverses the ranked list of documents from top to bottom, the user behavior can be described as follows: user presumably clicks on relevant documents to view and learn more about the topic, then moves on to the next relevant document. In order to maximize user gain, each relevant result must provide some new information that was not provided by previous relevant results. In other words, if the first document contained information about *techniques to build a wooden fence*, the user likely benefits more if the second document contains information about *materials required to build a fence*, rather than information about *techniques to build a wooden fence*.

The two search scenarios discussed above illustrate two extreme cases of the *novelty and diversity* problem. The first example highlights the problem of *ambiguity*

in the query, while the second focuses on the need for *coverage*. These queries act as exemplars for the two types of diversity identified in the literature: *extrinsic* and *intrinsic* [121]. Extrinsic diversity addresses the uncertainty in an ambiguous query where the intent is unclear, which is best served by a ranking of documents, each of which covers a different possible underlying information need or intent. Intrinsic diversity can be described as diversification that focuses on reducing redundancy and providing novel information for an unambiguous but underspecified information need. The query *how to build a wooden fence* discussed in the second scenario is a good example for intrinsic diversity.

Scenario 3

Lastly, we show that even queries that focus on the need for *coverage* could be ambiguous to some extent. In order to understand this, let us consider the same example query as in Scenario 2: *how to build a wooden fence*. Table 1.1 shows some possible distinct, smaller needs or *subtopics* for a broader topic about building wooden fences. We might hypothesize two different users with very different goals regarding the general topic: user A might be a middle school student writing an essay on the topic "wooden fences", while user B might be a do-it-yourself enthusiast trying fence the backyard. Therefore, user A's profile for the example query consists of subtopics a and c, and user B's of b,c and d. While the query *how to build a wooden fence* seemed to be well specified at first look, it is now ambiguous (or underspecified) by taking into account the task that initiated information need.

	subtopic	*user A*	*user B*
a.	Find basic information about building a wooden fence.		✓
b.	Find detailed information about building chain-link fences.	✓	✓
c.	What materials are best for building fences?	✓	
d.	Where can I buy materials for building a fence?	✓	

Table 1.1: An example topic (*how to build a wooden fence*) along with its subtopics and two possible user profiles indicating the interests of different users.

Therefore, the above discussion suggests that there is some amount of ambiguity present in any given query, either due to the ambiguity in the query terms or due to the user having underspecified their information need. An IR system must try to understand the information need underlying a query and, as best it can, provide the user with various pieces of information relevant to their need

We compare the whole process of searching for information to the question-answering task; the query posed to the retrieval system could be compared to a question and the documents retrieved could be compared to an answer to that question. Note that the question inherits the ambiguity in the query, and the subtopics are responsible for dealing with this ambiguity.

A visual representation of intrinsic and extrinsic diversity as modeled by subtopics can be given by grouping subtopics in a hierarchical tree-like manner in which higher levels account for ambiguity and the lower levels account for providing specific details to answer the query. For example, consider the query *nlp*; the first level would contain *neuro-linguistic programming* and *natural language processing* and other possible disambiguations as subtopics. Under *natural language processing* could be subtopics such as *natural language processing books*, *natural language processing techniques*, etc. Figure 1.2 gives an illustration of this hierarchical structure. We believe that for any query if we could provide the smallest set of documents containing all the subtopics answering the question, the overall user experience. And, we take into account that an IR system is used by a set of users with similar yet different information needs for the same query.

1.2 Preference Based Evaluation

The hierarchical tree-like representation of information need, although intuitive, is very hard to construct in practice for a given query (even a tree of depth one). Even clear and unambiguous queries can lead to different representations of information need for the given query due to the underlying task or user population. Furthermore, such a representation makes it difficult to incorporate other factors such as readability,

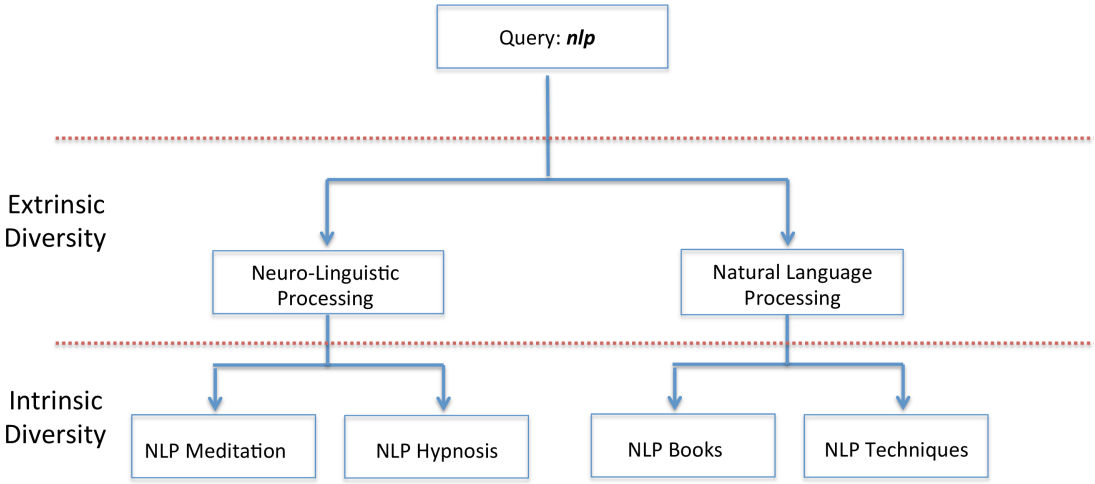

Figure 1.2: A hierarchical representing information need for an example query "nlp".

presentation, and other factors that play a role in determining the user's satisfaction. Therefore, we introduce a preference based framework to replace the idea of subtopics that is capable of capturing relevance and novelty and also naturally (and implicitly) incorporates those properties that influence the user's preference for one document over another.

Eliminating subtopics represents a complete overhaul in how researchers currently think about novelty and diversity. It requires a completely new framework for evaluation, including new metrics and new test collections. Much of this work is devoted to building this new framework and validating its assumptions and feasibility using both simulation and real user data.

1.3 Contributions

The following is a summary of our primary contributions in this work:

- **Developing Retrieval Systems**
  - We introduce a probabilistic set-based framework that maximizes the likelihood of covering all possible subtopics for a given query. The algorithm identifies the smallest set of documents that cover as many pieces of information as possible to satisfy the diverse information needs of the user.
- **Understanding User Preferences**
  - We study the factors that influence user preferences in novelty and diversity rankings. A user study was conducted to investigate how the presence of subtopics in a document influences user preference in the context of novelty and diversity. Simulations determined how well user-expressed preferences could match the information contained in subtopics.
- **Evaluating Retrieval Systems**
  - A statistical method to compare and analyze various evaluation metrics for novelty and diversity.
  - A conditional preference framework to estimate the utility of a document in a ranking. The framework has the potential to account for various factors implicitly, which enables better estimation of effectiveness.
  - A set of utility based metrics to measure the amount of novelty and diversity in a ranked list.

1.4 Organization

We discuss in detail the various components required to study the novelty and diversity problem. Initially the focus is on the subtopic framework that decomposes the information need into a set of subtopics (Chapter 3); the chapter introduces a probabilistic set-based approach to solve the novelty problem. Evaluation based on the subtopic framework is analyzed and compared to user's preferences by conducting a user study (Chapter 4). Based on the outcome of the study, we propose a novel preference based framework and develop a set of metrics for evaluating automatic retrieval systems (Chapter 5). Finally, we describe a large-scale user study that puts our methods to the test in a common IR evaluation scenario (Chapter 6).

Specifically:

- **Chapter 2 - Background Information** provides an overview of information need and briefly discusses various definitions of relevance from various viewpoints. The chapter also provides a survey of various retrieval models and evaluation methodologies proposed in the field of information retrieval, with a particular focus on novelty and diversity.
- **Chapter 3 - Models for Novelty and Diversity** introduces a probabilistic set-based approach that maximizes novelty and diversity in a ranked list. This chapter discusses various methods to identify and estimate subtopics for a given query, and a greedy approach to reduce redundancy.
- **Chapter 4 - Meta-Evaluation of Novelty and Diversity Evaluation** analyzes the subtopic based framework evaluation measures using statistical tools. The basic principles of the subtopic framework are identified and evaluated against user preferences with the help of a triplet of documents.
- **Chapter 5 - Novelty Evaluation using User Preferences** proposes a novel conditional preference framework to measure the utility of a document with emphasis on novelty. Then, we develop a set of preference based metrics that make use of the document utility to evaluate retrieval systems for our novelty and diversity task. The conditional preference based approach is compared to the traditional subtopic based approach by means of simulation.
- **Chapter 6 - Measuring System Effectiveness using User Preferences** presents a large-scale user study combining the ideas in previous chapters to evaluate both intrinsic and extrinsic diversity simultaneously in a common IR evaluation scenario.

- **Chapter 7 - Conclusion and Future Work** concludes the work with a look towards future directions in developing and evaluating retrieval models for novelty and diversity.

Chapter 2

BACKGROUND INFORMATION

Information retrieval (IR) is the process of obtaining relevant information for a given information need, which is a complex concept that has been of great interest to researchers for decades. The *information need* is considered to be the starting point of all searches, but it can be challenging to express as a question or query [16]. In general, information needs can be classified into two types: one in which users know exactly what they want, and therefore can describe their needs clearly to an information system (referred to as known-item search); and one in which a user only has a vague idea of the answer they are looking for, and tends to explore on the topic searched [63]. In the latter type, when a user's knowledge about their search topic is limited, query formulation is more difficult; the result is queries that are underspecified and ambiguous [14, 16]. Therefore, understanding the information need is vital in providing users with relevant information.

Relevance is probably the most important and most debated the concept in Information Sciences (IS) and IR [107]. Broadly, relevance has been classified into two types: "objective" or systems-based relevance and "subjective" or user-based relevance [159]. In truth, relevance is always subjective, in that it is entirely dependent on the context of the user and the task the user wishes to complete. However, most laboratory based IR experiments rely on a systems-oriented notion of relevance, according to which the relevance of a document is unchanging, dependent only on a representation of the document and a representation of the user's information need, and ignoring all external factors. In particular, this notion of relevance ignores the possibility that the relevance of a document in a ranking may depend on other documents in the same ranking.

2.1 Novelty and Diversity

The assumption that a document's relevance is independent of the relevance of other documents has been central to several theoretical and practical retrieval models in the past, most notably the *Probability Ranking Principle* [130]. But concerns of document overlap or redundancy was discussed as early as 1964 by Goffman [73], who recognized the need to address inter-document interactions in a ranked list and tried to incorporate them in his probabilistic retrieval model. The point was reiterated by Cleverdon [61], Brooks [22] and later by Boyce [21]; according to them, once a user examined the first document in a ranking, the document interaction effects will inevitable affect the choice of documents thereafter. Subsequently, Eisenberg et al. [71] conducted a study to provide some empirical evidence in support of incorporating document interaction effects into the evaluation process. A recent study that investigated the effect of showing highly relevant documents early in the assessment process found similar results [161].

The *novelty and diversity* problem arose as an attempt to model these interdependencies or interactions, which are so important to how users interact with search results. There have been several community-wide efforts to capture the idea of novelty and diversity experimentally. The TREC Novelty track, which ran from 2002–2004 [81, 164, 166], investigated the novelty problem within a broad topic. Systems were required to identify sentences within documents; those sentences should be both relevant as well as non-redundant with previous sentences from chronologically ordered stream. The Topic Detection and Tracking (TDT) workshops (1998–2004) [5] studied the "first story detection" problem that dealt with identifying previously unknown news event from a stream of chronologically ordered documents. Along the same lines, the TREC Question Answering track (1999–2004) [181, 180, 176, 178, 183] required systems to find answers to natural language questions such as "list 8 oil producing states in the United States". Systems assembled answer-strings representing facets or "nuggets" relevant to the question from multiple documents.

The common theme in these efforts is the attempt to model interactions among

relevant units of information (sentences, documents, nuggets) — when one provides the user with relevant information, it is generally not necessary to show the user another that provides the same information. But the common assumption is that there is a well-defined information need or "topic" against which relevance can be judged. As we discussed above, this is not always the case; queries are often underspecified or otherwise ambiguous, leading to differences of opinion in what should be considered "relevant".

The issue of query ambiguity was addressed by Fairthorne [72] in 1963. Despite that, IR evaluation evolved following the tradition of using detailed well-specified statements of an information need, with relevance assessed independently of other documents. This is partly because it is much simpler to implement this in experimental test collections. However, Clough et al. [62] analyzed query logs of a commercial search engine and estimated about 9.5% to 16.2% of all queries to be strongly ambiguous, while it is agreed that many more queries are underspecified to some degree [68]. Furthermore, users' browsing behavior as evidenced in query logs suggests their dislike to redundant documents in a ranked list [67, 50]. Recently, Sparck Jones et al. [170] urged evaluation methodologies to address query ambiguity, while Sanderson et al. [151] called for creation of test collections that account for multiple interpretation of a query.

An earlier effort to create topics with multiple notions of relevance was the TREC Interactive track (2002), which built a collection in which the relevance judgments accounted for multiple possible interpretations or *intents* of a query [114]. For example, given a query query *robotics*, assessors were required to identify a list of application relating to robotics such as *controlling inventory*, *spot-welding robotics*, *clean room*, etc. Expert judges identified query intents by reading retrieved documents and provided a short phrase describing the intents (Radlinski et al. argue that this approach lacks soundness and cannot capture the completeness of the information needs found in real scenarios [125]). The issue heated up with several works addressing the need for novelty and diversity in web search [2, 123, 58, 195]; the problem was further highlighted and discussed in a workshop at SIGIR, which identified different meanings

of the term "diversity" [121]. As a result, an effort to create intents that better reflect real user needs was realized by using query logs and served as the basis for TREC Web track's Diversity ranking task, which started in 2009 and continues to the present.

To us, the two sets of related work described above point to two separate, though related, problems: one is presenting a user with information that is both relevant to their information need but not redundant with information they have already seen; the other is presenting multiple users with the same query but different underlying needs with information that is maximally likely to be useful to all of them.

2.2 Retrieval Models

The primary goal of an IR system it to retrieve information that is relevant to the user's information need. Traditionally, IR research has approached this problem by returning a ranked list of documents in decreasing order of relevance as deemed by the system. Robertson formalized this approach in 1977 [130] as the *Probability Ranking Principle*:

> "If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data."

In the most common scenario, the retrieval models takes a document and query representation as input, and output a score indicating the probability of relevance for a document to the query. Retrieval models often represent queries and documents as bags of words to estimate the similarity between user request (query) and the document, ignoring their linguistic structure. Despite the fact that valuable information could be lost using the bag of words approach, it is has been shown to provide a reasonable first approximation [149].

Several retrieval models have been proposed with a pursuit of retrieving relevant information, under the assumption that information needs conveyed by a user in the

form a query is unambiguous. The Boolean retrieval model was one of the earliest of them; according to this model a query is represented as a Boolean expression of terms, and documents are represented as binary vectors indicating the presence of terms. A set of relevant documents is obtained by selected those documents that satisfy the Boolean expression. There is no notion of ranking documents; a document is either retrieved or not retrieved depending on the truth value of the Boolean query.

Following this, a *vector space model* was by Salton introduced to handle the importance of terms that occur in a query or document by using a multi-dimensional vector representation and provide a way to rank documents [148]. This allowed terms to be weighted by the frequency of occurrence in a document (*term frequency* or *tf*) and the inverse of the number of documents that contain the term (*inverse document frequency* or *idf*) to identify term that help discriminate relevant from non-relevant documents [132]. Depending on the specific values of *tf* and *idf* and the way they are combined, documents can be ranked according to how similar they are to the query by computing the cosine of the angle between a vector representing the document and a vector representing the query.

The Probability Ranking Principle, which formalized Salton's notion of ranking, resulted in the development of various models using probability theory. BM25 is by far the most popular model; it gives an approximation of a 2-Poisson model of term "eliteness" using a handcrafted equation that makes use of term frequency and inverse document frequency components to model relevance [135, 131]. The relevance score of a document D for query Q is given by

$$score(Q,D) = \sum_{w \in Q \cap D} tf_{w,Q} \frac{(k_1+1)tf_{w,D}}{k_1((1-b)+b\frac{|D|}{avg\_doclen}} log \frac{N - df_w + 0.5}{df_w + 0.5} \tag{2.1}$$

where the sum is over terms that occur in both query and document, $tf_{w,Q}$ is the number of times term w occurs in the query, $tf_{w,D}$ is the number of times term w occurs in the document, $|D|$ is the length of the document (i.e. the number of terms in it), avg_doclen is the average length of documents in the collection, N is the number of documents in

the collection, df_w is the number of documents term w occurs in the collection, and k_1 and b are free parameters.

More recently, probabilistic language models, a statistical technique that found much success in speech processing, were successfully adopted to IR [120, 137]. Since then, language models have grown in popularity, with several different versions appearing in the literature [168, 98, 18, 196]. The most common is known as the *Dirichlet-smoothed query-likelihood* model, which has the following form:

$$score(Q, D) = \sum_{w \in Q} \frac{tf_{w,D} + \mu \cdot \frac{ctf_w}{|C|}}{|D| + \mu} \tag{2.2}$$

Instead of inverse document frequency, language models use *collection term frequency* (ctf), the number of times a term appears in all documents in the collection, divided by the total number of terms in the collection ($|C|$). The language model has been particularly successful in that it can more easily do things like incorporating term dependencies by weighing query terms to improve retrieval performance [118, 105, 17]. Most models that have been developed for novelty and diversity build on the techniques discussed above or make use of them indirectly.

The retrieval models discussed above were developed for so-called *ad hoc* retrieval, in which a system must be able to take an arbitrary query and estimate relevance under the assumption that the query conveys the users' information needs unambiguously. These models estimate the probability of relevance based upon a single representation of user's information need, and further assume relevance of a document to be independent of other documents; the only things important to the scoring functions are the query representation and the single document representation. New retrieval models are needed to address the challenges of novelty and diversity discussed in Section 2.1.

2.2.1 Maximal Marginal Relevance and Reducing Redundancy

One of the first retrieval models to address the issue of redundancy in a ranked list was Carbonell and Goldstien's *Maximal Marginal Relevance* [30]. Given an initial

set of ranked documents (provided by one of the approaches above, for example), the MMR algorithm iteratively re-ranks them to produce a ranking with less redundancy. MMR uses a formula that combines relevance and "novelty", with novelty represented as lower similarity between two documents, in a linear manner, allowing the degree of novelty to be controlled using a parameter λ:

$$MMR = \arg\max_{D_i \in R; D_i \notin S} \left[\lambda \; score(D_i, Q) - (1 - \lambda) \max_{D_j \in S} (sim(D_i, D_j)) \right] \tag{2.3}$$

where R is an initial set of documents which is to be diversified, S is the current subset of document selected for the new ranking, $score(D_i, Q)$ is any scoring function which measures query relevance (e.g., BM25, language model, etc), $sim(D_i, D_j)$ is any similarity function which measure the similarity between two documents (cosine similarity is a commonly used measure), and λ is a parameter used to vary the amount of novelty. By using the original relevance score and discounting documents that are more similar to some already-ranked document, the algorithm prefers those documents that are both relevant to the query and also less similar to any document ranked in a prior iteration.

A probabilistic interpretation of MMR was given by Zhai et al. [197], in which probability of relevance $P(rel|D_i)$ and a probability of containing novel information $P(new|D_i)$ is combined into a scoring function conditional on previously-ranked documents $D_1, D_2, .., D_{i-1}$:

$$\begin{aligned} score(q, D_i | D_1, ..., D_{i-1}) = & c_1 P(rel|D_i) P(new|D_i) \\ & + c_2 P(rel|D_i) P(\neg new|D_i) \\ & + c_3 P(\neg rel|D_i) P(new|D_i) \\ & + c_4 P(\neg rel|D_i) P(\neg new|D_i) \end{aligned}$$

Zhai et al. argue that there is no cost to presenting a novel relevant document (i.e. $c_1 = 0$) and that the cost of presenting a non relevant document is unaffected by

whether that document is novel or not (i.e. $c_3 = c_4$), resulting in the final rank-equivalent scoring function:

$$score(q, D_i|D_1, ..., D_{i-1}) = P(rel|D_i)\left(1 - \frac{c_3}{c_2} - P(new|D_i)\right)$$

The ratio c_3/c_2 can be replaced with a single parameter ρ, and the problem reduces to estimating $P(rel|D_i)$ (which can be done with the query-likelihood score Eq. 2.2) and $P(new|D_i)$. Zhai et al. propose an estimate they call *AvgMix*, which is essentially the query-likelihood score with the query replaced by document D_i and the document replaced by a concatenation of all documents ranked up to position $i - 1$.

Similar to MMR, Carterette and Chandar [38] used a greedy approach that starts with a ranked list by relevance score and then prunes documents that are similar to higher-ranked documents beyond some threshold, while Chen and Karger [51] proposed to optimize the probability of obtaining at least k relevant documents in a given set of retrieved documents.

2.2.2 Models Based on Subtopics

In Section 2.1, we noted the ideas of *nuggets* in the TREC Question Answering track and *intents* behind ambiguous queries in the TREC Interactive track. The idea behind both is the same—to decompose the query into simpler units that documents can be judged more precisely. Since the word "topic" has come to be used in IR as a representation of an information need, these simpler units have come to be known as *subtopics*. The idea of retrieval models based on subtopics is that the best way to improve a user's experience with a search engine is to try to identify subtopics rather than retrieving broadly relevant documents. Therefore, their goal is to retrieve a ranked list of documents such that each one contains novel subtopics that are either relevant to the information need or relevant to some alternative intent.

Subtopic-based models typically follow a standard framework: given an initial ranking by relevance score, an automatic retrieval system hypothesizes a set of subtopics, scores documents against these subtopics, then re-ranks them with a diversity ranking function. An example illustrating a typical subtopic-based approach

is given in Figure 2.1. Given a query such as *nlp*, an initial ranking of documents is obtained using a relevance function such as BM25 (Section 2.2). Then, the system hypothesizes a set of subtopics for *nlp* and estimates the probability that a document is relevant to each of those subtopic; this produces the matrix of probabilities shown. The subtopics identified by the system in the figure include *neuro-linguistic programming*, *natural language processing* and *nlp basketball*. Finally, a diversification function is used to produce a diverse ranked list. In our example, ideally, the system must produce the following ranked list — $\{Doc_1, Doc_2, Doc_4, Doc_3\}$; the ranked list covers all three subtopics higher in the ranking satisfying users with different intents early.

Such models have two primary components: a module for hypothesizing subtopics (often using data mining techniques) and a module for re-ranking documents. First, we discuss the various approaches to identifying subtopics; we follow that by summarizing various diversity ranking functions proposed in the literature.

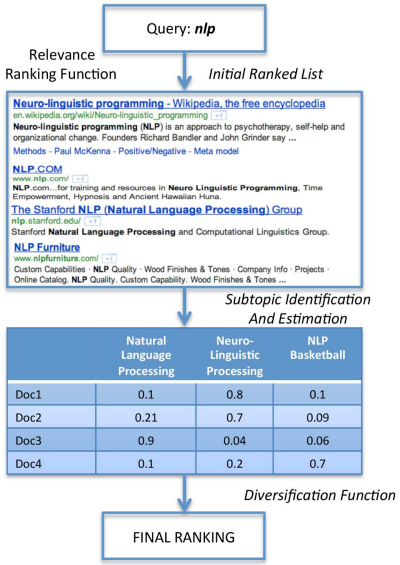

Figure 2.1: An example illustrating the workflow and various component in a subtopic based approach to model novelty and diversity.

2.2.2.1 Subtopic Mining Techniques

Methods proposed to identify a set of subtopics for a given query can be classified into two categories: *implicit* and *explicit*. Implicit methods model subtopics using information contained within documents obtained from an initial retrieval, whereas explicit methods make use of external information such as query log, Wikipedia disambiguation, links between documents, etc, to identify subtopics for a given query. Table 2.1 categories various subtopic mining techniques proposed in the literature.

Table 2.1: Representative subtopic mining approaches in the literature, categorized by the source of information the model requires to generate subtopics.

Implicit Methods	**Explicit Methods**
Topic Models [38]	Query Suggestions [155]
Cluster Analysis [38, 84]	Clicks and Reformulation [124]
Language Models [197, 38]	Taxonomy-based [2]
Pattern Mining [199, 190]	Wikipedia [88]
Entity-based [38]	Webgraphs [43]

Many of the implicit methods represent subtopics as a bag of words; in particular methods based on topic modeling or clustering represent a subtopic as a weighted collection of words that may have no obvious meaning to a human. Nevertheless, such approaches have been illustrated to improve diversity ranking. Several approaches can be broadly thought of as "clustering": k-means clustering of documents; hierarchical clustering; clustering salient phrases from search result snippets [190]; hierarchically clustering snippets according to themes; discovering common phrases; identifying meaningful groupings using Latent Semantic Indexing. Documents obtained from different high-quality clusters are more likely to be relevant and diverse; thus, a set of high-quality clusters may represent the diverse information needs for a query. Typically, documents contain relevant and non-relevant information; the non-relevant information in a document could lead to noisy clusters. A pattern-based approach was introduced to overcome this limitation by identifying semantic relation between term through their

co-occurrences in retrieved documents [199]. Another approach considered representing various hypothesized subtopics for a query by constructing of relevance models or topic models [38].

Explicit approaches leverage external sources such query logs, Wikipedia disambiguations, Open Directory Project, or other sources to model subtopics for a given query. Many of these represent subtopics as short phrases that would be understandable to humans. Santos et al. [156] exploited query suggestions (or related queries) from popular web search engines to improve diversification effectiveness. Several methods use query logs, including reformulation, session, and query-URL clicks data [191, 101, 29]. Strohmaier et al. [171] employ a random walk similarity algorithm to group queries obtained from similar session, the frequency of quires in the log can be used to estimate popularity (or importance). These methods suffer the data sparsity issue as a given query might not be present in the taxonomy or logs might not have enough user data to extract subtopics. Furthermore, there are various settings where query logs or any external information is not available.

Once a set of subtopics has been obtained, it is usually straightforward to compute a score for each document and each subtopic: regardless of whether it is represented as a bag of words or a short phrase, the subtopic can be treated as a query in any of the scoring functions detailed in Section 2.2. This produces the matrix of document-subtopic scores illustrated in Figure 2.1.

2.2.2.2 Diversity Ranking Functions

Once a set of subtopics and scores have been obtained, the final step is to obtain a ranking that maximizes subtopic coverage and minimizes redundancy. Various functions have been shown to improve system performance; we discuss them below. The general idea behind all of them can be expressed mathematically as follows:

$$DivFun(Q, D, R) = \lambda\ rel(D|Q) + (1 - \lambda) div(Q, D, S, R) \tag{2.4}$$

where Q is the user's query, D is the document being scored for diversity, R is the original ranking of documents by relevance score, and S is the set of subtopics. $rel(D|Q)$ is

the relevance score for document D (again, this could be any function in Section 2.2), and $div(Q, D, S, R)$ is a diversity score computed from the query, the document, the original ranking, and the subtopics. The λ parameter is used to tune the tradeoff between relevance and diversity in the final ranking.

Since the relevance score can (in theory) be any of those discussed in Section 2.2, we differentiate the models below by their implementation of the diversity function.

WUME – Yin et al. [194] proposed a diversity function that combines the relevance and coverage of subtopics given a query. The function that they refer to as *WUME* considers only the subtopic coverage of a document ignoring novelty. The diversification function is shown below:

$$div_{WUME} = \sum_{s \in S} P(s|Q)P(D|s) \tag{2.5}$$

where $P(s|Q)$ gives the probability that a user cares about subtopic s, with a higher probability indicating greater importance, and $P(D|s)$ is the score based on the content of the document and subtopic, weighted by the likelihood that the document belongs to a particular subtopic. The function maximizes the coverage of subtopics present in a ranked list.

IA-Select – Agrawal et al. [2] studied the problem of answering ambiguous web queries by classifying information in both queries and documents. They present an approach to increase user satisfaction by taking into account the various intents a user have for a query. A query q could belong to a set of categories $C(q)$ and a document d could belong to a set of categories $C(d)$. Assuming that the probability distribution of categories for the query $P(s|q)$ is know, finding a set of documents with diverse ranking is possible by maximizing the function

$$div_{IA-Select} = \sum_{s \in S} P(s|Q)(1 - \prod_{d \in R}(1 - P(d|s))) \tag{2.6}$$

where $P(s|Q)$ gives the probability that a user cares about subtopic s, with a higher probability indicating greater importance, and the second part of the equation ($\prod_{d \in R}(1-P(d|s))$) iterates over previously ranked documents R penalizing documents with redundant subtopics to promote novelty. The function considers those subtopic that are not covered in the ranked list as more useful.

xQuad – Santos et al. [156] proposed a probabilistic function that maximizes the relevance of of a document to the given query and the novelty of each subtopic given the previous ranked documents in the ranking. The coverage and novelty components are combined into a single entity in this function. The function is a greedy algorithm that picks a document that maximizes the below function to add to the final diverse ranking.

$$div_{xQuad} = \sum_{s \in S} P(s|Q)P(D|s) \prod_{d \in R}(1 - P(d|s)) \tag{2.7}$$

where $P(s|Q)$ gives the probability that a user cares about subtopic s, with a higher probability indicating greater importance, and $P(D|s)$ is the score based on the content of the document and subtopic, weighted by the likelihood that the document belongs to a particular subtopic. Final part of the equation $\prod_{d \in R}(1 - P(d|s))$ estimates the novelty of a document D by comparing it to previously ranked documents R for each subtopic s.

Square Loss Function – A diversification function similar to the xQuad, but it uses a square loss function to account for the novelty a document provides to the ranking. The coverage and novelty components are combined into a single entity in this function as well [200]. The function computes the novelty of each subtopic for all the documents and picks the document that maximizes the function given below:

$$div_{SqLoss} = \sum_{s \in S} P(s|Q)P(D|s)(2 - 2\sum_{d \in R}(P(D|s) - P(D|R)) \tag{2.8}$$

where $P(s|Q)$ gives the probability that a user cares about subtopic s, with a higher probability indicating greater importance, and $P(D|s)$ is the score based on the content

of the document and subtopic, weighted by the likelihood that the document belongs to a particular subtopic. The $(2 - 2\sum_{d \in R}(P(D|s) - P(D|R))$ part of the formula is a square loss function that takes into account previously ranked documents to capture novelty.

For all of these models, λ would be determined from training data; the same value of λ would be applied to every query. However, it is not necessarily the case that each query should be diversified equally; some could benefit from selective diversification. Santos et al. [158] proposed to learn the diversification tradeoff parameter on a per-query basis. They gathered a pool of query features and learn a model in order to determine the diversification tradeoff for unseen queries, thereby improving overall diversity ranking effectiveness.

2.2.3 Other Approaches to Diversity Ranking

MMR (and variants) and approaches based on subtopics have been the primary models developed in the IR literature. There are other approaches as well. Online learning method that exploit users' feedback to learn diversity or novelty function have been considered; Radlinski et al. [123] proposed an online learning algorithm that optimized on the number of clicks found in the logs for a given query to improve effectiveness. Researchers have also proposed offline supervised learning scenario to utilize various machine learning technique for the problem [103, 122, 157, 162]. Yue et al. [195] constructed a training data using initially retrieved documents and subtopic level judgments to learn a function that maximizes coverage. A similar supervised learning approach within an online setting was presented by Raman et al. [126].

Building upon the foundations of quantum mechanics, Zuccon and Azzopardi [202] introduced a novel quantum probability principle that takes into account the relevance of other documents in the ranked list. Other approaches to the diversification problem includes: a risk minimization framework based on portfolio theory [189]; modeling diversity of a query directly at the term-level [69]; an election-based approach [70];

an axiomatic framework for result diversification [74]; a method to combine subtopics from multiple sources [84]. More recently, data fusion techniques for diversification were proposed and found to positively impact system performance [102].

2.3 Evaluation

Information retrieval evaluation aims to predict the retrieval system's ability to satisfy user's information need. The question of what and how to of IR evaluation is tied to the retrieval task at hand, which in turn dictates all other experimental design choices such as the notion of relevance, etc. A typical IR task is *ad-hoc retrieval*: a user enters a keyword query, and the retrieval system returns a ranked list of documents. Within the ad-hoc retrieval task researcher identified different sub-tasks such as known-item retrieval [15] and topic distillation [66] where users are looking for more specific information, passage retrieval task in which the unit of retrieval is a passage [85, 112]. Other tasks commonly studied within the field include: information filtering or routing [134]; automatic detection and tracking of emerging stories in stream of text (Topic Detection and Tracking) [5]; searching medical scholarly articles for various gene names [86]; searching semi-structured documents focusing on retrieval of document fragments [97]. While our work might be relevant to other tasks, we restrict our study to a task similar to *ad-hoc retrieval*, where the information need underlying a query is diverse, and the system is expected to return a ranked list of documents containing relevant and non-redundant information.

Two types of IR evaluation has been identified in the past: user-based and system-based evaluation [182]. The goal of a user-based evaluation is to measure system effectiveness taking into account various factors that affect user satisfaction in real time. A user-based evaluation tries to create a scenario as close as possible to a scenario faced by real users using an IR system, and aims to estimate system performance. On the other hand, system-based evaluation simulates a scenario to reflect an interaction between a real user and a retrieval system. The simulation consists of a set of information needs (topic), and a set of relevance judgments for a set of document for each

topic. A set of information needs are created by assessors (or task organizers), along with a keyword query. In general, the information needs are expected to represent the actual needs of the user. Eventually, assessor judge documents returned by systems determining their relevance for a given information need. Evaluation measures use these relevance judgments to predict system effectiveness. System-based evaluation of less expensive compared to user-based evaluation as the relevance judgments can be reused to test new systems developed.

In early 1960s, the work of Cleverdon and his collaborators to build the Cranfield collections is considered as the starting point of system-based IR evaluation. Cranfield, a small test collection was constructed to test retrieval systems in a laboratory setting. The collection consisted of 1,400 research articles written on aerodynamics [60], 221 queries, and relevance judgments for every document for each query. The queries and relevance judgments were obtained by sending out questionnaires to authors asking them to write a natural language question summarizing the problem addressed in their paper; these became the collection topics. (For a detail historical review of IR evaluation, see Chapter 1 in [152]) The methodology later came to be known as the *Cranfield Methodology* and has become the standard in information retrieval evaluation [187].

According to the Cranfield style of evaluation, systems under comparison produce a ranked list for a set of topics and effectiveness scores are averaged. The methodology makes three simplifying assumptions as pointed out by Voorhees et al. [182]: (1) only topicality of the document is considered to estimate relevance, (2) a single assessor can capture the need of the entire user population, and (3) for each topic all possible relevant documents in the collection are known. Various strategies have been proposed in the past to deal with scenarios when some assumptions are violated. A great deal of research effort has gone into studying the problem, and they show that it is possible to evaluate systems effectively even when the assumptions are violated. The rest of this section gives a brief overview of the standard experimental design for system based evaluation of *ad-hoc retrieval* task and extends the discussion to our diversity problem when applicable.

2.3.1 Test Collection

A test collection consists of three distinct components: a collection of documents (often known as *document corpora*), a set of information needs that are represented by statements (*topics*) or a set of keywords (*queries*), and a set of relevance judgments. The relevance judgments are a list of relevant documents for each topic that are expected be retrieved by the retrieval system. We discuss each of these components in detail below.

Document Corpora

Document corpora consist of a set of retrievable units (often documents) from which an IR system searches for relevant information. The choice of the corpora is determined by the retrieval task [82, 75, 111] and the availability of the corpora [188]. Creating a test collection is a difficult task [172], the corpora must contain documents on a variety of topics while making sure enough documents are present in the collection for a single topic. If the documents are on several different topics, it is much easier to create queries for the test collection. Care must be taken to ensure that each topic does contain enough relevant documents in the corpora. Finding the right balance is important, as an estimation of system effectiveness over a set of diverse topics would result in more generalizable conclusions. Also, the document collection must enable the creation of information needs that are representative of user needs in the real world.

The early days of IR evaluation depended on a series of smaller test collection such as ADI, MEDLARS, OSHMED, etc. [150, 100, 87] that often consisted of a collection of research literature or news articles. Sparck-Jones and Van Rijsbergen stressed on the need for larger collections and proposed their idea of an *ideal test collection* [169]. Their vision of creating a large test collection was the foundation for a community-based evaluation standard (Text REtrieval Conference – TREC). In 1992, TREC was started with the initial goal of building a large test collection to evaluate TIPSTER (a text retrieval research project) [188]. Thenceforth, the evaluation program has operated on an annual cycle with the goal of creating test collections for various retrieval tasks. The created test collections are made available to the research community every

year. Over the years, the size of the corpora has progressed from a few thousand to millions of documents (the first TREC collection), to billions of webpages and other documents.

Construction of document corpora for tasks such as web page retrieval is possible by obtaining a subset of available webpages on the internet. Crawlers are employed to collect a set of webpages; they begin with a certain number of seed webpages, extracting the content and out-links present in each page, and proceed in an iterative manner. In other words, the crawlers simply sample a subset of the webpages on the world wide web. The sample size and sample bias can affect the accuracy and quality of the retrieval systems considerably [83]. Thus, different crawling strategies can be engaged to bias the crawler towards desired pages based on some criterion such as: link-based popularity [52] , topicality [42], user interests [115], and avoidance of spam [77]. We use a newswire corpus created by Allan et al. [6] and a web corpus (Clueweb 09) distributed as part of the Lemur Project [53] (for details, please refer to Appendix A)

Topics

A set of topics need to be created as part of the test collection. The systems are evaluated based on their performance on these topics. Since, the goal of system-based evaluation is to estimate the real world performance of a system, chosen topics must be realistic and closer to information needs of real users using the system. Further, each topic must contain enough relevant documents in the document collection. Thorne [175], in his work suggested sampling queries from logs to create a testing environment closer to a real world scenario. Although, in practice log of user queries are not available for many document collections.

Often, topics are created by annotators who provide a detailed account of their information need along with a short keyword query. Usually, the same annotator who developed the topic is hired to assess the relevance of documents retrieved. Topics for traditional collection including TREC are developed in this manner. Topics with too few or too many relevant documents are omitted to avoid biases towards very common and very rare documents respectively [186, 80]. Section 2.4.1.1 deals with the topic

creation for the novelty and diversity task that accounts for ambiguity in the query by requiring a set facets to satisfy an information need. Although, a general theory of developing good topics for evaluation remains an open question for both *ad-hoc* and *novelty & diversity* tasks.

2.3.2 Relevance Judgments

Accurate and reliable measurements of relevance is fundamental to the evaluation of IR systems [95]. IR evaluation has traditionally focused on topical relevance: a document is considered relevant even if the document contains a single phrase or sentence on the topic of interest. Typically, relevance judgments are obtained by showing the information need along with a document to a human annotator. The annotator examines the document to determine its relevance to the information need. Also, the annotator needs to quantify the degree of relevance in a document indicating how useful the document for a given information need.

The relevance judgments are in turn used by evaluation measures to estimate system effectiveness. Traditionally, IR evaluation has adopted to the use a binary scale in which a document is either relevant to a topic or not. Retrieval tasks such as web search and other precision oriented tasks require finer distinction between documents, relevance on a 3- or 5-point has become a standard in the industry for web search evaluation [192, 40, 89].

A document's relevance for a given topic is almost always determined independent of other documents, know an *absolute judgments*. Rorvig [136] proposed a method to assess relevance by comparing two document side-by-side, known as *preference judgments*. Studies suggest that *preference judgments* can often be made faster than graded judgments, with better agreement between assessors (and more consistency with individual assessors) while making much finer distinctions between documents [129, 37].

For several publicly available test collections, relevance judgments were created by community-based evaluation efforts such as TREC, NTCIR, INEX [97, 93, 187]. The organizers hire a group of well-trained annotators to develop topics and judge a

set of documents for a given topic to determine their relevance. While this style works well for ad-hoc retrieval task, it becomes challenging when judgements from multiple annotators need to be incorporated. An alternate approach of using crowdsourced workers was described in the work by Alonso et al. [8]. In their work, workers were hired using an online labor marketplace known as Amazon Mechanical Turk and the relevance judgments were obtained from these workers. Such platforms enable us to recruit a diverse pool of annotators making it possible to capture the diverse information needs, which is desirable for our work.

Assessor Disagreement

Typically, the relevance of documents is assessed by a single assessor; however, relevance is known to differ across assessors [177]. Even for the same assessor relevance changes with time [160]. In 2000, Voorhees [179] constructed a test collection with relevance assessments from additional assessors to study the impact of assessor disagreement on evaluation. On average, the study observed an agreement of only 42%, i.e. two assessors agreed on the relevance of a document only 42% of the time. However, these large variations in relevance judgments did not seem to affect the relative performance of retrieval systems. It must be noted that when alternate assessments are obtained, additional assessors attempt to understand the information needs by looking a query (or statement of information) in order to make the judgments. This process could certainly be a factor leading to disagreement amongst users. The novelty and diversity problem argues that the source of disagreement on relevance of a document is partly due to their differing information needs and ambiguity in the query.

2.3.3 Evaluation Measures

Evaluation measures estimate the effectiveness of a system by computing a score for each ranked list returned by the system for a set of queries. The score is proportional to the total relevance of a ranked list for a given query. In the previous section, we discussed various ways to estimate a document's utility (relevance of document). The next step is to capture the cumulative relevance (or usefulness) of a ranked list.

Historically, the metrics were aimed at estimating the effectiveness of a set of documents. Measures such as precision, recall, and F-Measure were proposed; these measure aimed at determining if a document should be retrieved or not. Precision is the proportion of retrieved documents that are relevant whereas recall is the proportion of relevant documents that were retrieved. A measure that takes both these factors into account is the F-measure; defined as the weighted harmonic mean of recall and precision.

The advent of the digital era along with the information overload problem meant that more relevant documents were available for a given query; therefore documents were examined sequentially from top to bottom. Thus, a browsing model in which a user scans a rank list from top to bottom, one document at a time and stops at some rank became the norm. Precision and recall values were computed at rank cut-off to reflect such a browsing model. Average-precision (AP) is perhaps the most widely used metric and is computed by averaging the precision scores at each relevant document in a ranked list [80]. Mean Average Precision (MAP), the mean of average precision values for a set of topics became a common way of reporting system performance. Jarvelin and Kekalainen proposed the *nDCG* measure to model a user who prefers to see highly relevant documents at top ranks [91, 90]. They relied on graded relevance judgments to model document utility, although their approach can be used with binary judgments as well. Discounted Cumulative Gain at rank k for a given query is defined as:

$$DCG@k = \sum_{i=1}^{k} \frac{2^{g_i} - 1}{log(i+1)} \tag{2.9}$$

where g_i is the relevance grade of the document at rank i. The metric rewards documents with large relevance grades, while discounting the gains at lower ranks. Thus, the measure favors systems that rank highly relevant documents high in the ranked, which is the foundation for most precision-based metrics such as Ranked Biased Precision (RBP) metrics [110]. RBP can be defined as

$$RBP = (1-p) \sum_{i=1}^{k} g_i p^{i-1} \tag{2.10}$$

where g_i is the relevance grade of the document at rank i and p represents the probability that the user will persist looking through the ranked list, and $1-p$ gives the stopping probability. More recently, Chapelle et al. [50] introduced Expected Reciprocal Rank (ERR) that takes into account the relevance of previously ranked documents. According to ERR, the probability of relevance of a document diminishes as the relevance of higher ranked documents increases and is defined as:

$$ERR@k = \sum_{i=1}^{k} \frac{1}{r} \prod_{r=1}^{i-1} (1 - p_r) p_i \tag{2.11}$$

where p_i denotes the probability that the document at rank i is relevant. In practice, p_i is defined as a function of the relevance grade of a document, i.e. $p_i = (2^{g_i} - 1)/2^{g_m ax - 1}$ with the maximum possible grade for a document denoted by g^{max}. Other rank-biased metrics that have been reported in the literature include: reciprocal rank (the reciprocal of the rank of the first relevant document retrieved) [94]; R-precision [10]. The measures presented above are the most common for evaluating ad hoc retrieval.

2.3.4 Incompleteness in Relevance Judgments

As test collections became larger and larger, the problem of assessors not being able to judge every document in the collection had to be addressed. To deal with this problem, a common strategy known as *pooling* was employed. Pooling is a document selection strategy where only the union of top k documents retrieved by each system per topic are judged. Pooling makes the judgments incomplete thus violating the assumption made by the Cranfield paradigm, but studies have been conducted to validate the use of pooling [79, 201]. The methods used in these studies encouraged researchers to experiment with shallower pooling depths. Later, it was found that even a shall pooling depth of 5 produces a good approximation to evaluations done with a pool depth of 100 [35].

Typically, IR researchers are faced with evaluating a set of retrieval system. For instance, TREC style community evaluation needs to obtain relevance judgments to evaluate a set of submitted systems. The goal, then, is to compare the effectiveness

of these submitted systems. Hence, efficient document selection strategies could be employed to judge documents that help discriminate systems by their performance. Cormack et al. [65] proposed a technique called "Move-to-Front Pooling" that prioritizes documents by the rank and system that retrieved the document. A similar method was introduced by Zobel et al. [201] in which documents from a shallow pool is judged first and then, the second pool is constructed based on the knowledge obtained about the systems and topics in the initial pool. Other algorithmic approaches of selecting documents include: selection of documents to best approximate full ranking of systems [109] or comparing two systems [41], a hedge algorithm using an online setting [11]. Although the algorithmic approaches reduce the number of relevance judgments required, the values of evaluation metrics estimated by these approaches are hard to interpret.

Therefore, new approaches and measures were proposed to deal with incomplete relevance judgments. The proposed methods include: the *bpref* measure that counts the number of document pairs that were swapped when compared to an ideal ordering [24], inferred average precision (*infAP*) [193], condensed list (where unjudged documents are excluded from the ranking) [141], making relevance prediction by assuming relevance of a document to be uniformly distributed [36].

For the novelty and diversity task, a shallow pool of 10 or 20 is used to obtain a pool of documents. While there exist several studies to validate the use of pooling to obtain an unbiased sample of documents for the ad-hoc evaluation, only simple pooling techniques are being used for the diversity task and its validation is still an open question to be considered in the future.

2.3.5 Reliability of Evaluation

Ideally, system-based evaluation must predict the expected effectiveness of a retrieval system in real search scenarios. However, the methodology used to construct test collection significantly affects the prediction quality. Nevertheless, system-based

evaluation fosters rapid research and development by providing an easy and accessible testing framework.

System-based evaluation must strive to restrict the source variance to the retrieval algorithm being tested [133]. Topic sample is a major source of variance; thus several studies in the literature have been dedicated to investigate this effect [184]. Using an analysis of variance model, Banks et al. [13] showed the topic effects to be the largest source of variability. The work pointed out that the choice of topics influence system performance considerably and the way a system treats each topic contributes to its effectiveness significantly. To deal with this variability, effectiveness scores are typically average over a set of topics. A natural question that arises is, how many topics are needed to distinguish system effectively. Empirical studies suggest that statistical difference in effectiveness scores over 50 topics [185, 23] with relative score difference of more than 10% are considered reliable [154]. There has been several studies conducted to study the effect of pooling depth and topic size with the goal of providing a set of guidelines leading to more reliable evaluation [184].

2.3.6 Repeatability and Reusability

The test collection are expected to facilitate both repeatability and reusability. Repeatability enables researchers to produce the same result every time the experiments are conducted under the same environment. Whereas, reusability ensures that relevance judgments in a test collection are complete allowing researchers to compare any new methods that they develop in the future. Comparison of old systems against a new system is possible without requiring any additional relevance judgments in such a setting.

The use of pooling described earlier makes relevance judgments incomplete and thereby raising concerns about reusability of a test collection. Zobel et al. [201] conducted a study in which they held out one or a set of runs submitted by a group to TREC for construction of document pools. They observed that missing relevant documents did not seem to affect the relative performance of the systems. More recently,

a method to quantify the reusability of a test collection was proposed by Carterette et al. [39]. Reusability of test collection for novelty and diversity requires investigation into how unjudged documents affect effectiveness of a system [144] as well as how coverage of subtopics or nuggets (that represent diverse information needs) affects system performance.

2.3.7 Analysis of Evaluation Measures

Evaluation measures play a vital role in analyzing the performance of a system, comparing two or more systems, and optimizing systems to perform some task. Evaluation measures models the user's satisfaction with a retrieval system by a single score, which is a complex process. The complexity involved along with different ways to model user satisfaction has resulted in several different evaluation measures. Analyzing the strengths and weaknesses of commonly used evaluation measures is essential for improving and understanding them. Tague-Sutcliffe and Blustein [173] were amongst the first to compare different evaluation measures. They analyzed the correlation between system rankings as determined by different precision-based evaluation measures. The correlations were useful in identifying if the evaluation metrics measure similar or different aspects of a system.

IR system developers and researchers often rely on an evaluation metric to answer questions like: "Does system A significantly outperform system B?". Naturally, the ability of the metric to discriminate between systems becomes a concern. Zobel et al. [201] proposed a *swap method* to test the predictive power of a measure. According to their method, topics were split into two halves, then a pair of runs is compared using an evaluation measure using the first half and later using the second half. If the order of pairs remained the same across the two halves, the measured result is considered correct or else the result is considered a mistake. Similar studies were done on different test collections to compare P@10 and MAP by their predictive power [23, 185, 165, 154]. A measure of co-variance (Cronbach's Alpha) [20], and a bootstrap test to count statistical significance between pairs of runs [142] are other notable ways to measure stability.

However, it is worth mentioning that a measure that sorts systems by *runid* will be identified as perfectly stable by the above methods [152].

Alternatively, researchers in the past have tried to measure the correlation between user preferences and predictions made using a test collection and evaluation metrics. A study conducted by Al-Maskari et al. [3] looked for correlations between the effectiveness scores returned by different evaluation measures and user satisfaction. Rank-biased measures such as DCG were found to show strong correlating with user measures. Similar observations were made by other researchers as well [89, 4]. A study by Smith and Kautor [163] engaged users to search using two different versions of a web search engine. One version displayed results from rank 1 whereas the second displayed results starting from a much lower rank (both rankings were obtained from the same IR system). The latter system was presumably much worse, although to their surprise, there was no significant difference in user satisfaction. The authors reported that the users adapted to the poorer system by issuing more queries (often reformulating queries) to deal with the smaller number of relevant documents retrieved. The dynamic nature of the users make accurate measurement of user satisfaction extremely challenging.

2.4 Novelty and Diversity Evaluation

The evaluation for the novelty and diversity task must account for multiple information needs underlying a given query (*diversity*) and the amount of novel information in a ranked list (*novelty*). An intuitive way to model diversity and novelty is to use a set of subtopics to represent a given query. Each subtopic is expected to reflect a specific information need of a query, and the presence of a subtopic in a document indicates the relevance towards that information need. The subtopics enable modeling of novelty by examining the number of previously unseen subtopics in a ranked list. We explain the construction of test collection (Section 2.4.1) using the subtopic approach and also discuss various evaluation metrics (Section 2.4.2).

2.4.1 Test Collections for Novelty and Diversity

Typically, a test collection consists of a collection of documents, a set of information needs (or topics), and judgments of relevance for each document per topic. If subtopics are used to model novelty and diversity, then a set of subtopics reflecting user's information needs for a query has to be curated. Additionally, at least two levels of relevance judgments are needed: binary or graded judgments indicating the relevance of each document, and for each relevant document, a list of subtopics contained within the document.

The creation of topics in a test collection is nontrivial for any IR task; the need to develop topics with multiple information needs makes it even more challenging. Also, the test collection must contain a representative set of subtopics for each topic. There exists two popular ways of obtaining those subtopics: (1) for each relevant document; annotators are required to judge the relevance of a document against a pre-defined set of subtopics, and (2) the annotators create subtopics upon examining relevant documents by either highlighting phrases or creating their own subtopic labels. Section 2.4.1.1 discusses these methods in detail. Once, a set of topics (along with subtopics) are available, the next step is to obtain relevance judgments for each document. A pool of documents are selected for a topic, then human annotators are asked to review the relevance of each document. Typically, a pool depth of 20 is used to select a set of documents that are judged by trained assessors. Generally, a binary grade is assigned to each document indicating the presence of each subtopic for a given query.

Two test collections suitable for studying the problem of *novelty and diversity* in English that are publicly available include: the TREC Web Track Diversity Task dataset [53] and the *newswire data* prepared by Allan et al. [6]. We use both of them, please refer to Appendix A for a detailed description of the datasets. The NTCIR Intent task dataset is another test collection similar to the TREC dataset is used to studying this problem in Chinese and Japanese [93].

2.4.1.1 Topic Creation

In this section, we briefly talk about two popular methods of identifying a representative set of subtopics for evaluation purposes. We believe an exhaustive enumeration of all subtopics for a query is very hard, thus rely on alternative way to obtain subtopics implicitly using preferences (discussed in Chapter 4 and 5). We argue that different methods used for subtopic creation could lead to considerable differences in evaluation.

Query Logs – The logs of commercial search engines often keeps track of user's interactions such as queries issued to the system with timestamp, reformulation, document clicked for a query, etc. Often such logs contain millions of queries that provide a rich source of information reflecting user's real information needs. Evidence from clicks and reformulation provides information about the relatedness between queries; for example, if a document is clicked often for two different queries then the queries are more likely to be related. The TREC Web track Diversity task dataset is created using query log with the help of a tool that clusters query reformulations of the given query obtained from query logs [124, 53]. Clusters obtained using such methods were manually examined to filter strange and unusual interpretations.

Document Annotation – Alternatively, subtopics can be obtained by assessors examining documents to identify relevant text fragments (a phrase or paragraph) and group them by defining labels that represent subtopics. Notice unlike the TREC dataset the subtopics were not provided to the assessors beforehand. The *Newswire* dataset compiled by Allan et al. [6] is created this way, it consists of three levels of judgments for each query: binary relevance judgments, list of subtopics for each relevant document and a passage in the document supporting its relevance to the subtopic. Recently, a similar nugget-based approach was used to identify subtopics [117, 116, 145].

The differences between the two methods are subtle, yet it leads to different number of subtopics being created for each topic having practical implication. The

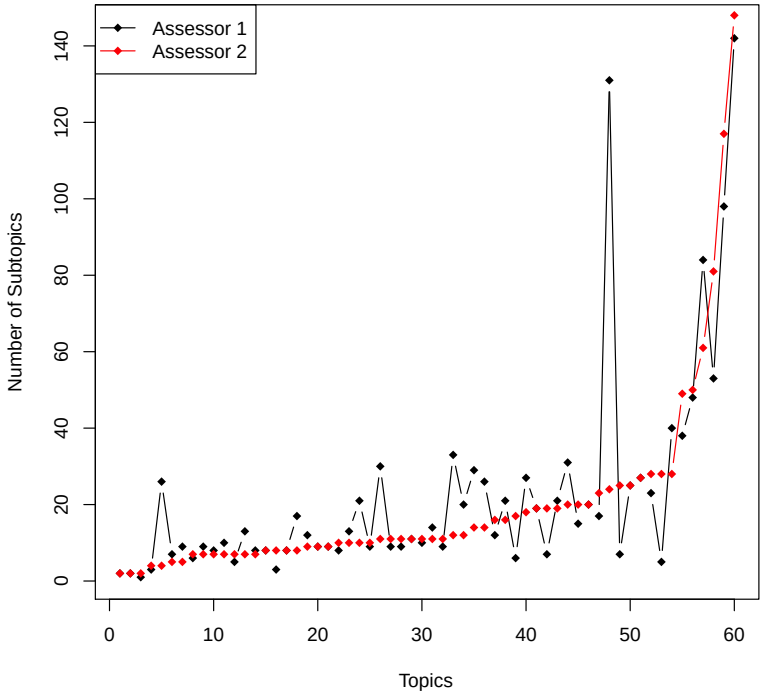

Figure 2.2: Number of subtopic identified by the two annotators for each query for the newswire dataset.

number of subtopics for the TREC queries ranges between two and eight with majority of the queries containing three to four subtopics. The *Newswire* contains two set of judgments obtained from two different annotators. Figure 2.2 shows the disagreement on the number of subtopic between the annotators. This clearly demonstrates that there is substantial disagreement between assessors about the number of subtopics. While the document annotation approach generates more subtopics, the coverage of these subtopics is still questionable due to the limitation on the number of documents examined. Secondly, although search engine logs contain interaction from millions or users, the subtopic identification algorithm induces a bias towards popular intents when data is sparse. Also, the subtopics obtained using these methods are often broad and less suitable for novelty evaluation.

2.4.2 Evaluation Measures for Novelty and Diversity

Evaluation measures for novelty and diversity must account for both relevance and novelty in the result set. It is important that redundancy caused by documents containing previously retrieved subtopics are penalized while documents containing novel information are rewarded. Most evaluation measures solve this problem by requiring that the subtopics for a query be known and that documents have been judged with respect to subtopics. An alternate to the subtopic approach was proposed by Chandar and Carterette [48]; they proposed a set of utility based metrics to calculate the effectiveness of the rank list. The work proposed herein discusses these utility-based metrics in detail.

The subtopic approach has been used predominantly for reporting effectiveness scores for novelty and diversity. A simple set-based measure called *subtopic recall* was introduced by Zhai et at. [197] to estimate the coverage of subtopics in a ranked list. Subtopic recall does not take into account the importance of subtopic nor does it account for novelty. To overcome this limitation, the intent-aware metrics extended the traditional ad-hoc measure to estimate effectiveness in the presence of multiple

subtopics for a query [2]. Clarke et al. [58] introduced α-nDCG, a cascade-based metric to encourage the coverage of multiple subtopic while penalizing the presence of redundant subtopics in a ranked list. Other measures include ERR-IA [49], NRBP [57], D-Measures [143], etc. We explain in detail various measures used in this work.

Subtopic Recall – measures the number of unique subtopics retrieved at a given rank [197]. Given that a query q has m subtopics, the subtopic recall at rank k is given by the ratio of number of unique subtopics contained by the subset of documents up to rank k to the total number of subtopics m.

$$S\text{-}recall@k = \frac{\left|\bigcup_{i=1}^{k} subtopics(d_i)\right|}{m} \tag{2.12}$$

α-nDCG – scores a result set by rewarding newly found subtopics and penalizing redundant subtopics. In order to calculate α-nDCG we must first compute the gain vector [58]. The gain vector is computed by summing over subtopics appearing in the document at rank k:

$$G[i] = \sum_{j=1}^{m} (1-\alpha)^{c_{j,i}-1} \tag{2.13}$$

where $c_{j,i}$ is the number of times subtopic j has appeared in documents up to (and including) rank i. Once the gain vector is computed, a discount is applied at each rank to penalize documents as the rank decreases. The most commonly used discount function is the $log_2(1+i)$, although other discount functions are possible. The *discounted cumulative gain* is given by

$$\alpha DCG@k = \sum_{i=1}^{k} \frac{G[i]}{log_2(1+i)} \tag{2.14}$$

α-DCG must be normalized to compare the scores against various topics. This is done by finding an "ideal" ranking that maximizes α-DCG, which can be done using a greedy algorithm. The ratio of α-DCG to that ideal gives α-nDCG.

Intent-aware Family – Agarwal et al. [2] studied the problem of answering ambiguous web queries, which is similar to the subtopic retrieval problem. The focus of their

evaluation measure is to measure the coverage of each intent separately for each query and combine them with a probability distribution of the user intents. They call this the *intent-aware* family of measures. It can be used with most of the traditional measures for evaluations such as precision@k, MAP, nDCG, and so on.

ERR-IA Expected Reciprocal Rank (ERR) is a measure based on "diminishing returns" for relevant documents [50]. According to this measure, the contribution of each document is based on the relevance of documents ranked above it. The discount function is therefore not just dependent on the rank but also on relevance of previously ranked documents. A weighted average of the ERR measures for each interpretation would give the intent-aware version of ERR [49].

D-Measures – The D and the D# measures described by Sakai et al. [143] aims to combine two properties into a single evaluation measure. The first property is to retrieve documents covering as many intents as possible and second is to rank documents relevant to more popular intents higher than documents relevant to less popular intents.

Table 2.2 shows a toy example with the subtopic judgment matrix for two different systems: one more diverse than the other. Table 2.3 shows the evaluation scores for the two systems. The score are computed at rank cut-off 5. All four measures agree with ranking of the two systems, *i.e.* System X diversifies more than System Y, although the difference in scores between the two system is least for Precision-IA. We analyzing these and more aspects of the above described measure in Chapter 4.

2.4.3 Analysis of Novelty and Diversity Evaluation Measures

Earlier in Section 2.3.7, we briefly discussed the prior works to evaluate the ad-hoc evaluation measures and methodologies. Here, we discuss the efforts taken to evaluate the novelty and diversity evaluation methodology. Novelty and Diversity evaluation is still a new topic; there are few studies that evaluate the evaluation methodology. To compare the ability of a measure to distinguish between systems,

System X

documents	subtopics a	b	c	d	e
d_1	✓			✓	
d_2		✓			
d_3		✓			
d_4					
d_5	✓				✓
d_6	✓				
d_7			✓		
d_8	✓				
d_9					
d_{10}					

System Y

documents	subtopics a	b	c	d	e
d_2		✓			
d_3		✓			
d_4					
d_6	✓				
d_8	✓				
d_5	✓				✓
d_1	✓			✓	
d_7			✓		
d_9					
d_{10}					

Table 2.2: A toy example with 8 documents and 5 subtopics. The first ranked list is visible more diverse than the second

Run	ERR-IA	$\alpha - nDCG$@5	Prec-IA@5	s-recall@5
System X	0.3970	0.7707	0.2400	0.8000
System Y	0.2179	0.4187	0.1200	0.4000

Table 2.3: Effectiveness scores for the two toy example systems in Table 2.2 returned by ERR-IA,$\alpha - nDCG$, Precision-IA and subtopic-recall at rank 5.

Clarke et al. [54] studied the properties of novelty based evaluation measure by comparing the measures in terms of discriminative power and rank correlation. Sakai et al. [146] compared some of the evaluation measures discussed in Section 2.4.2 using *discriminative power*. The method involves pairwise computation of significance test between the runs for a given measure. And the number of significant pairs reflects the discriminative power of the measure.

Carterette [33] studied the mathematical properties of the diversity measure, namely subtopic-recall and α-nDCG showing that computation of an ideal rank list in these measures is an NP-Complete problem. His work showed that even for a small number of subtopics a greedy algorithm could overestimate the minimum rank or ideal gain vector required by s-recall and α-nDCG respectively, which introduces marginal errors into calculations of these measures. Chapelle et al. [49] showed that ERR-IA

and α-nDCG are members of a family of metrics called Intent-Aware Cascade-Based Metrics. The work provides a theoretical analysis of the intent aware measure and shows that the measures exhibit sub-modular properties. Chandar and Carterette [44] compared the subtopic-based measure such as α-nDCG, ERR-IA and MAP-IA using TREC Web track Diversity task data. They isolated different effects such as diversity, relevance and ranking using ANOVA analysis for each measure.

Section 2.3.7 provided a brief overview of various studies intended to compare user satisfaction and test collection based evaluation for ad-hoc retrieval. Here, we outline the efforts taken towards validating novelty and diversity measures against user satisfaction. Sanderson et al.[153] used the TREC Web track Diversity task dataset to study the predictive power of ad-hoc and diversity metrics. They compared a pair of ranked lists and measured the agreement between user preference and prediction made by an evaluation measure for diversity, and report that intent recall is as effective as more complex metrics such as *alpha*-nDCG. Chandar and Carterette [47, 46] described the subtopic-based measures such as $\alpha - nDCG$, ERR-IA, S-Recall using a set of principles and test how well these principles hold true with respect to user preferences.

Chapter 3

MODELS FOR NOVELTY AND DIVERSITY

Traditionally, models of information retrieval are built under the assumption that the relevance of a document is independent of other documents in the ranking: two identical documents are both relevant as long as they contain information the user needs. Modeling documents as independently relevant does not necessarily provide an optimal user experience. Certainly, five relevant documents that all contain the same single piece of information are not as useful to a user as one relevant document that contains five separate pieces of information – yet all the traditional evaluation measures described in Section 2.3.3 would reward a system that provides the former more than one that provides the latter. This problem could be resolved by modeling documents as *inter-dependent* rather than independent — if a system could use the fact that its top five documents are identical to one another, it would know to rank only one of them. The notion of subtopics can be used to model inter-document dependencies: two documents relevant to the same subtopic are considered 100% inter-dependent, while two documents relevant to two different subtopics are independent.

In this chapter, we focus our attention on developing methods that hypothesize subtopics for a given topic, and rank documents with respect to these subtopics with the goal of reducing redundancy while maintaining topical relevance. The goal, then, is to identify a small set of documents that cover as many unique subtopics as possible.

3.1 Retrieval Models

IR systems operate with the goal of helping a person find useful information in response to a need. We describe two challenges faced by such systems as follows: (1) different users with different information needs might issue the same query, but the

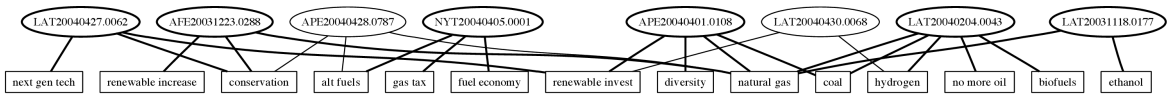

Figure 3.1: Example document-subtopic graph. An edge from a document to a subtopic indicates that the document attests the subtopic. The bolded document nodes indicate the smallest set needed to cover all of the subtopics.

system has no knowledge of the differing needs. Thus, systems are expected to deal with these diverse information needs in order to satisfy a group of users. (2) often several pieces of information (or subtopics) are required to satisfy an information need of a single user and these pieces of information may be spread across several documents. As we have discussed, the first case is called *extrinsic diversity* and is quite common in web search. The second is often referred to as *intrinsic diversity* and may occur whenever a query is underspecified. In this section, we propose to solve the latter challenge using a probabilistic set-based approach.

Section 2.2 provided a brief overview of the existing methods for novelty and diversity problem. Of those models, the Maximal Marginal Relevance [30] and Zhai et al. [197]'s approaches are most suitable to our specific task. They work under the assumption that similar documents cover similar subtopics and, therefore, consider the subtopics of an underlying query implicitly.

Our model consists of the following four major components: hypothesizing a set of subtopics, estimating probabilities that a document is relevant to a subtopic, and a diversity ranking function that combines these scores to obtain a final diverse ranking. We detail our probabilistic set-based approach in Section 3.1.1, discuss various methods to hypothesize a set of subtopics in Section 3.1.2, and the diversity functions are explained in Section 3.1.3.

3.1.1 Probabilistic Set-Based Approach

We argued that an information need can be decomposed into a set of subtopics.

Each document can be relevant to one or more subtopics, and each subtopic can be represented in one or more documents. Thus documents are modeled as subsets of the subtopic set, and subtopic relevance is modeled as a document 'containing' a subtopic.

Figure 3.1 shows relationship between document and subtopics using a bipartite graph for an example query *reduce dependence oil*. The underlying information need for the query is:

> Many countries are trying to reduce their dependence on foreign oil. What strategies have countries, organizations, or individuals proposed or implemented to reduce the demand for oil? This could include exploring new sources of energy or reducing overall energy demand.

The subtopics of this need include *invest in next generation technologies, increase use of renewable energy sources, invest in renewable energy sources, double ethanol in gas supply, shift to biodiesel, shift to coal,* and more. In Figure 3.1, 14 subtopics that are relevant to the information need are shown along with eight relevant documents. Each edge reflects the containment of a subtopic in a particular document.

Our set-based formulation suggests a set-based ranking principle: *i.e.* to retrieve a set of documents that maximize the likelihood of capturing all of the subtopics. This can be visualized by generalizing Figure 3.1 to a graph in which instead of 0-1 edges between documents and subtopics, each edge has a weight representing the probability that each document contains every possible subtopic. A probabilistic interpretation of the graph is one in which every document has some probability of containing every subtopic, and the thickness of the edge reflects the strength of the belief. Thus, the idea behind the probabilistic set-based approach is to find the smallest set of documents with maximum probability of containing all the subtopics.

Suppose we have a query q representing the user's information need and set of documents D retrieved by a relevance ranking function. Suppose our system has identified a hypothetical set of subtopics S of size m. The goal is to estimate the probability that S is contained in D *i.e.* $P(S \in D)$. Assuming that a subtopic occurs in a document independently, we can define the probability of a particular subtopic S_j in a document set D where D has documents $\{D_1, D_2\}$ as

$$P(S_j \in \{D_1, D_2\}) = P(S_j \in D_1 \ \cup \ S_j \in D_2)$$
$$= P(S_j \in D_1) \ + \ P(S_j \in D_2) - P(S_j \in D_1, S_j \in D_2)$$

We assume that a subtopic occurs in a document independently, so we have

$$P(S_j \in \{D_1, D_2\}) = P(S_j \in D_1) \ + \ P(S_j \in D_2) - (P(S_j \in D_1) \cdot P(S_j \in D_2)) \quad (3.1)$$

Alternatively, the above equation can be written as follows [1] :

$$P(S_j \in \{D_1, D_2\}) = 1 - (1 - P(S_j \in D_1))(1 - P(S_j \in D_2)) \quad (3.2)$$

In general, then, the probability that a subtopic S_j occurs in at least one document in a set D where D has documents $\{D_1, D_2, ...D_n\}$ is

$$P(S_j \in D) = 1 - (1 - P(S_j \in D_1))(1 - P(S_j \in D_2))...(1 - P(S_j \in D_n)) \quad (3.3)$$

$$= 1 - \prod_{i}^{n}(1 - P(S_j \in D_i)) \quad (3.4)$$

and the probability that a set of documents contains all the subtopics in the set S is given by

$$P(S \in D) = \prod_{j=1}^{m} P(S_j \in D) = \prod_{j=1}^{m} 1 - \prod_{i}^{n}(1 - P(S_j \in D_i)) \quad (3.5)$$

Maximizing $P(S \in D)$ with respect to a document set D should result in a diverse ranking which covers all the subtopics in the set S.

There are four main components in our set-based framework:

0. A relevance ranking for the original query needs to be obtained before using our set-based framework.
1. Find a set of subtopics for a given query (information need), i.e. identify the set of subtopics S.
2. Estimate the probability that a subtopic is contained in a document, i.e. estimate $P(S_j \in D_i)$ for each $S_j \in S$ and $D_i \in D$.
3. Rank documents such that the top k (subset D_k) is likely to contain all possible subtopics.

[1] $P(A \ \cup \ B) = 1 - ((1 - P(A)) \ (1 - P(B)))$ when events A and B are independent.

3.1.2 Hypothesizing Subtopics

We describe three methods for finding a set of subtopics S and estimating $P(S_j \in D_i)$. Each method takes as input a query and a ranking of documents by relevance score. Each method outputs an $n \times m$ matrix of probabilities similar to that shown in Figure 2.1. The names or representations of the subtopics themselves are *not* part of the output, as they are not important to the final ranking.

3.1.2.1 Relevance Model - Subtopic Models

We propose a relevance modeling approach in which the retrieved set of documents for a given query is used to build subtopic models. A relevance model is a distribution of words $P(w|R)$ estimated from a set of relevant or retrieved documents R. Since, a relevance model can be estimated from a set of document, we use this approach to estimate m different subtopic models. The subtopic models $P(w|S_j)$ are estimated from the retrieved documents using the RM2 approach described by Lavrenko et al. [98].

$$P(w|S_j) \propto P(w) \prod_{s_k \in S_j} \sum_{D_i \in D_{S_j}} P(S_k|D_i) P(w|D_i) p(D_i)/P(w) \tag{3.6}$$

where D_{S_j} is the set of documents relevant to subtopic S_j, s_k are the subtopic terms, $P(w) = \sum_{D_i \in D_{S_j}} P(w|D_i)P(D_i)$, and $P(w|D_i)$ is a smoothed estimate. Since we do not know the subtopic terms or the set of documents relevant to the subtopic, we will estimate them from the retrieved documents.

We obtain m models from the top m retrieved documents by taking each document along with its k nearest neighbors as the basis for a subtopic model. Subtopic models can be built from a set of documents as done in ad-hoc retrieval, which can be seen as query expansion/relevance feedback methods (please refer [32] for details on query expansion and relevance feedback). However, instead of a single expanded query, there are m, where m is the hypothesized number of subtopics. Therefore, we obtain

$P(S_j \in D_i)$ for each retrieved document for every subtopic to obtain $n \times m$ matrix of probabilities. In this work we assume constant, manually-selected m for all queries.

3.1.2.2 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a state-of-the-art topic modeling approach for identifying topics in discrete datasets such as a text corpus. A topic model is a statistical method for identifying abstract *topics* that may be present in a collection of documents. Blei et al. [19] proposed LDA as a graphical model for topic discovery in which documents are a mixture of a small number of topics and each word in the document contributes to the topic. Given a set of documents, LDA identifies a set of topics and outputs a set of topic models with the probability that a word generates each topic.

Our hope is that two different topic models will capture something about the vocabulary associated with different subtopics by assigning higher probabilities to different terms. For example, we may have a subtopic model that corresponds to "biofuels" by giving higher probabilities to words that co-occur with "biofuel" more often than they co-occur with other terms. Another model that corresponds to "gas tax" by giving higher probabilities to words that co-occur with "gas tax" more often than other terms. Then $P(D_i|S_{biofuel}) > P(D_i|S_{gastax})$ suggests that document D_i is more likely to contain the "biofuels" subtopic than the "gas tax" subtopic (where $P(D_i|S_j) = \prod_{w \in D_i} P(w|S_j)$.

In LDA, probabilities $P(w|S_j)$ and $P(S_j)$ are found through expectation maximization. Then we find $P(D_i|S_j) = \prod P(w|S_j)$ and $P(S_j|D_i) \propto P(D_i|S_j)P(S_j)$, generally assuming a uniform prior on documents. Finally, we estimate $P(S_j|D_i)$ by applying LDA to a set of retrieved document using a parameter m that determines the number of topic models. In this work we assume constant, manually-selected m for all queries.

3.1.2.3 Webgraphs

The documents (or webpages) on the web are often linked to one another using hypertext, creating a link structure between documents. The link structure of documents has often provided a rich source of information about the content of the environment. We exploit this link structure to find several densely linked collection of hubs and authorities within a subset of the documents retrieved for a given query. Each densely linked collection could potentially cover different subtopics for a given query.

A set of documents retrieved for a given query consisting of hyperlinked webpages are represented as a directed graph $G = (V, E)$: nodes correspond to webpages, and a directed edge $(p, q) \in E$ to the presence of link from page p to q. We refer to a graph obtained by considering only a set of retrieved documents for a query as a sub-graph. The sub-graph is expanded to include all other documents in the collection linked to a document in the sub-graph (*in-links*) via hypertext. The *out-links* from the sub-graph, i.e. other documents to which the documents in the sub-graph links to are also added. The *hubs* and *authorities* scores are calculated for each document using the iterative procedure described by Kleinberg [96]. Kleinberg's procedure begins by representing the directed graph as an adjacency matrix. The adjacency matrix is multiplied by its transpose to calculate the principal and non-principal eigenvectors. Each element in the eigenvector corresponds to a document. The values in the principal eigenvector correspond to the *hub* score. The non-principal eigenvectors represent other densely linked clusters in the graph.

We select the first m eigenvectors, constructing a language model for each of the eigenvectors. Then, m language models are constructed from the documents correspond to the k greatest values in each of the first m eigenvectors. The intuition is that the link structure clusters the documents into subtopics; therefore, these models provide a hypothetical set of subtopic models. In all our approaches, documents are scored against the m hypothesized language models by considering it as a query expansion problem. Thus, m expanded queries with each query consisting of k top terms in the

language models are used to estimate $P(S_j \in D_i)$.

3.1.3 Diversity Ranking Function

A diversity ranking function takes the $n \times m$ matrix of $P(S_j \in D_i)$ probabilities produced by one of the three methods described above and outputs a diversified ranked list. We use two different approaches. One follows from the set-based formulation of the problem: we take the k documents that have highest probability of containing each of the m subtopics. Implementing this is simple; given the $n \times m$ matrix of probabilities produced by a method in Section 3.1.2, we can just find the maximum value in each column and take the corresponding document to be part of the retrieved set. We then rank these documents in decreasing order of their original relevance ranking score. We call this the *max-set* method.

The second approach is an iterative greedy algorithm based on Equation 3.5. The first step is to determine which document to rank first; to do that, we compute the following probability:

$$\begin{aligned} P(S|D) &= \prod_{j=1}^{m} 1 - (1 - P(S_j \in D_i)) \\ &= \prod_{j=1}^{m} P(S_j \in D_i) \end{aligned}$$

for each document D_i in our set D. In words, the first document is the one with the highest probability of containing subtopic S_1 AND subtopic S_2 AND subtopic S_3 and so on.

The second document will be the one that increases the total probability $P(S \in D)$ by the greatest amount *given* that the first document has already been placed. Let D_1 refer to that document. Then:

$$D_2 = \arg \max_{D_i; i \geq 2} \prod_{j=1}^{m} 1 - \prod_{i=1}^{n} (1 - P(S_j \in D_i))$$

The algorithm continues in that way, at each step fixing k documents in the ranking and identifying the one that most increases the total probability. Since each

step can be seen as computing a marginal likelihood (see Carterette and Chandar [38] for details), we call this the *marginal likelihood* method.

3.2 Experiments

We test and validate our proposed approaches against two different datasets: TREC Web track Diversity task dataset and Newswire dataset (see Appendix A for details). The primary set of experiments validates our probabilistic set-based approach for the *intrinsic diversity task*. Although our approaches are not strictly designed for it, we also test them against the TREC Web track Diversity task dataset.

3.2.1 Intrinsic Diversity Ranking

In this section, we describe various experimental retrieval systems for intrinsic diversity ranking task and validate them using data annotated with subtopics.

3.2.1.1 Data

We use the *Newswire* dataset created by Allan et al. [6] (for a detailed description refer to Appendix A). We use the 60 topics each annotated by two assessors; each topic consists of a short (3-6 word) query and relevance judgments for documents in a larger collection of newswire documents. There are three levels of judgment: a binary relevance judgment for the document; for each relevant document, a list of subtopics that the document contains; and for each subtopic, a passage in the document that supports its relevance to that subtopic. For each query, only the top 130 documents retrieved by a query-likelihood language model were judged. Since few documents were judged, it is very possible that subtopics that exist in the corpus do not appear in the judged documents. To ensure we have judgments on all ranked documents, we will only re-rank these 130 documents for each query.

3.2.1.2 Evaluation Measures

Zhai et al. [197] evaluated subtopic retrieval using a measure called S-recall and Zhang et al. [198] stressed on evaluating for redundancy separately. We use both the

measures to evaluate the experimental systems for our intrinsic diversity task.

S-Recall – The primary evaluation question for an intrinsic diversity ranking is how many of the subtopics that are identified in the corpus were retrieved by the system. Given a set of subtopics and documents judged according to whether they are relevant to the information need and contain each subtopic, we use the subtopic recall measure as defined in Section 2.4.2.

For each query, S-recall is computed at minimum rank at which perfect recall can be achieved. Because, the maximum value of S-recall at a particular rank k depends on the maximum number of subtopics that can be found in k documents. For the example in Figure 3.1, S-recall@1 can be at most 5/14 and S-recall@2 can be at most 8/14; at least 6 documents are required to achieve S-recall = 1. Here, 6 is the minimum rank at which perfect recall can be achieved, and we will denote S-recall at that rank simply S-rec. We argue that the best way to satisfy an intrinsic diversity need is to retrieve the smallest set of documents that contain all of the subtopics, and thus that S-rec is the most natural measure to evaluate a subtopic retrieval system.

Redundancy – When a subtopic S_j occurs in the document at rank 2 after having already occurred in the document at rank 1, its appearance in document 2 is redundant. There is often a tradeoff between eliminating redundancy and retrieving the smallest set of documents that contain all the subtopics: less redundancy may require more documents to cover all the subtopics. Therefore, we evaluate redundancy at rank k separately from recall and precision. Redundancy is the average number of times each subtopic is duplicated up to rank k (if there are no relevant documents ranked above rank k, redundancy is undefined). Between two systems with the same subtopic recall, the one with lower redundancy should generally be preferred, but lower redundancy is not by itself a reason to prefer a system.

3.2.1.3 Methods

In this section, we describe the different experimental systems for the intrinsic diversity task used in our experiments.

- LM baseline: a basic query-likelihood (Dirichlet smoothing; $\mu = 1000$) run with no subtopic model (refer to Section 2.2 for more details).
- RM baseline: a pseudo-feedback run with relevance modeling [98] and no subtopic model.
- MMR: maximal marginal relevance with query similarity scores from the LM baseline and cosine similarity for novelty. Query-likelihood scores are re-scaled to [0, 1] to make them compatible with cosine similarities. Section 2.2.1 provides a detailed description of this method.
- AvgMix: the probabilistic MMR model proposed by Zhai et al. [197] using query-likelihood scores from the LM baseline and the AvgMix novelty score as described in Section 2.2.1.
- SimPrune: a simple greedy approach that diversifies the result set by iterating through the initial ranking (LM baseline) and removing similar documents. The greedy pruning algorithm iterates over an initial ranked list sorted by relevance and prunes documents with similarity scores above a threshold θ. At *rank* i any document D_j is pruned if $j > i$ and $Sim(D_i, D_j) > \theta$.
- FM: the set-based subtopic model described in Section 3.1.2. We use two different ways to hypothesize subtopics and score documents.
  - FM-RM refers to the subtopic relevance model described in Section 3.1.2.1. Each of the top m documents and their k-nearest neighbors becomes a "subtopic model" $P(w|S_j)$ – a truncated (v-term) relevance model constructed from the documents. Then we compute the probability $P(D_i|S_j)$ for each document and subtopic model; these are converted to a probability $P(S_j \in D_i)$ by linear transformation to the range [0.25, 0.75]. A final ranking from $n \times m$ matrix of probabilities $(P(D_i|S_j))$ were obtained by using the marginal-likelihood approach described in Section 3.1.3.
  - FM-LDA uses subtopics discovered using the LDA method described in Section 3.1.2.2. This provides $p(z_j|D_i)$ for each document D_i and each "subtopic" z_j; these were used as the subtopic-document scores. The marginal-likelihood approach described in Section 3.1.3 used the subtopic-document scores to produce a final ranking. For each query, 50 subtopics were extracted.
- Manual: we use the actual labels in the Newswire data as subtopics in FM. Specifically, each of the subtopic labels is submitted as a query to produce the

matrix of probabilities, then we use the marginal-likelihood method to produce a final ranking. This is a "cheating" run (since normally the subtopic labels would not be known in advance) but could give a sense of the upper limits of automatic system effectiveness.

The *Newswire* (Section 3.2.1.1) data that we use to evaluate our experimental systems comprises of news articles, thus documents does not contain any hyperlinks. Since, no link structure is available, we do not include the Webgraph method (discussed in Section 3.1.2.3) to hypothesize subtopics.

We implemented all the models described above using the Lemur toolkit, as well as standard language modeling and language modeling plus pseudo-feedback with relevance models. Whenever possible, we have used the same similarity or scoring functions between models to ensure the fairest possible comparison. A five-fold cross-validation was used to train and test systems, and to obtain results for all 60 queries for each model. Specifically, we divided the 60 queries into five folds of 12 queries each. The 48 queries in four folds are used as a training set to select model parameters such as λ, θ, (m, k, v) (for MMR, SimPrune, and set models, respectively). These parameters are used to obtain ranked results on the remaining 12 queries. The query splits were chosen randomly in advance so that all experiments used the same training and testing data.

3.2.1.4 Results

For each method, we report S-recall at the minimum optimal rank subtopic-recall (which ranges from 0 to 1, larger values indicating better performance), redundancy at the minimum optimal rank (which has a minimum of zero but no upper bound; smaller is better), and mean average precision (MAP) using the document-level relevance judgments. Table 3.1 shows S-recall, redundancy ratio, and mean average precision (MAP) for all systems described above. Among the seven automatic methods, SimPrune gives the best overall results, though we note that there is no significant difference in the S-recalls of MMR, SimPrune, and FM-RM. All three retrieved about 44% of the subtopics, compared to roughly 40% by the two baselines and AvgMix,

Run	Subtopic-Recall	Redundancy	MAP
LM Baseline	0.405	0.856	0.583
RM Baseline	0.376	1.176	**0.617***
MMR	0.440	0.538	0.534
AvgMix	0.398	0.720	0.570
SimPrune	**0.444***	0.567	0.501
FM-RM	0.440	0.674	0.574
FM-LDA	0.153	**0.224***	0.285
manual	0.677	0.672	0.698

Table 3.1: S-recall and redundancy at the minimum optimal rank and average increase in S-recall from rank 1 to the minimum optimal rank for four subtopic topic retrieval systems. Numbers are averaged over 60 topics with two sets of assessments each. The best automatic result for each column is in bold. An asterisk indicates statistical significance.

factor	df	F	p-value
system	4	5.615	0.000
assessor	1	1.018	0.317
system:assessor	4	1.689	0.153

Table 3.2: Two-way ANOVA results on S-recall for the LM baseline, MMR, AvgMix, SimPrune, and FM-RM. Differences between systems are significant while differences between assessors do not significantly affect the results. There is insignificant interaction between assessor and system.

and only 15% by FM-LDA. All of the models (except FM-LDA) exhibited a fairly high degree of redundancy, duplicating each subtopic at least 0.5 times (on average) in the relevant documents retrieved by the minimum rank. MMR had the least redundancy, significantly lower than SimPrune and FM-RM. FM-LDA has very low redundancy, but this can be explained by the fact that it retrieved very few relevant documents.

The "manual" results in Table 3.1 provide some loose upper bounds. It retrieved 68% of the subtopics, but still retrieved each of them almost as many times as the subtopic model. This suggests that a fairly high degree of redundancy is inevitable. Many of the harder-to-find subtopics are only present in documents that contain easy-to-find subtopics; it is simply not possible to retrieve all of these without some redundancy. Therefore, the manual run suggests that lower redundancy is only

superficially desirable; optimizing for redundancy may result in "harder" subtopics being missed. This manual run also suggests that the set-based model is easily improved simply by improving the subtopic models: if a user provides some information about the subtopics, we can easily incorporate it into the subtopic models. This stands in contrast to MMR or AvgMix, which cannot incorporate such information as easily.

Since there were two assessors for each topic, we test hypotheses about differences between systems using a two-way within-subjects ANOVA on S-recall. A two-way ANOVA calculates the variance in a measurement of recall due to differences between systems and due to differences between assessors, as well as interactions between the two. We would like to see that the variance due to systems is significant and outweighs any other source of variance. If this is the case, the comparison is robust to differences in assessors. Ideally we would like to see that variance due to assessors is not significant, and in particular that the interaction is negligible. Table 2 shows a summary of the results of the ANOVA test to determine whether significance among the top five automatic runs is affected by assessor disagreement. Indeed, system differences are significant, while assessor differences are not, and there is negligible interaction between system and assessor.

Figure 3.2 shows the 11-point S-precision/recall curves for seven systems, it gives a better sense of how redundancy varies with S-recall for each run. The curves for MMR, SimPrune, and FM-RM curves are clearly above the LM baseline and AvgMix curves. Note that at both highest and lowest recall levels, the FM-RM curve is above the others but not in between. The FM-LDA system significantly underperforms compared to the others. The figure also shows redundancy increasing with S-recall. The LM baseline and AvgMix have the highest redundancy.. FM-LDA and SimPrune coincide closely over all S-recall values with FMRM in the middle. MMR and the manual run coincide closely up to S-recall 0.5; after that the redundancy of MMR increases to match or exceed that of SimPrune.

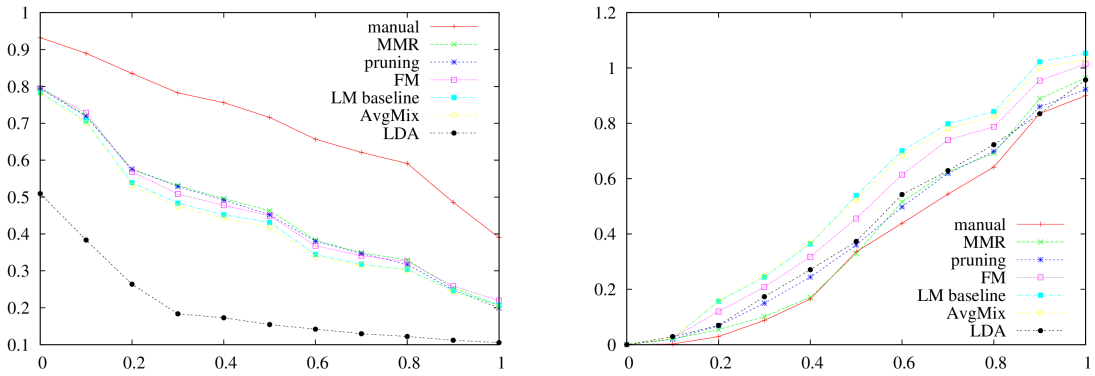

Figure 3.2: Subtopic-Recall vs. Subtopic-precision (left) and redundancy (right) for seven models

3.2.2 Extrinsic Diversity Ranking

Next, we test our methods for the extrinsic diversity task using the TREC Web track Diversity task dataset. The track uses a web dataset that consists of 50 million web pages in English. We use a total of 50 queries, and two level of relevance judgments: judgment on traditional relevance and subtopic level judgments. The Lemur Toolkit and the Indri search engine were used in our experiments. The query-likelihood result set with Dirichlet smoothing ($\mu = 2000$) (as described in Section 2.2) was used as our baseline for reranking.

For this data, we use the SimPrune and RM-FM models described above, and also a model we call RM-WebGraph that infers subtopics using links between web pages (see Section 3.1.2.3). The TREC data includes information about 454,975,638 links between 428,136,613 web pages, including the 50 million we used.

We evaluate our methods using two of the evaluation measures described in Section 2.4.2 which rewards novelty and diversity, namely α-normalized discounted cumulative gain (α-nDCG) and intent-aware precision (P-IA). All our methods were evaluated at rank 10 with α set to 0.5 for α-nDCG. We report the diversity results for our experimental runs along with the Indri baseline model in Table 3.3. The table shows that our methods perform poorly in diversifying the results set according to α-nDCG compared to the other proposed methods such as WUME and xQuad (Section 2.2.2).

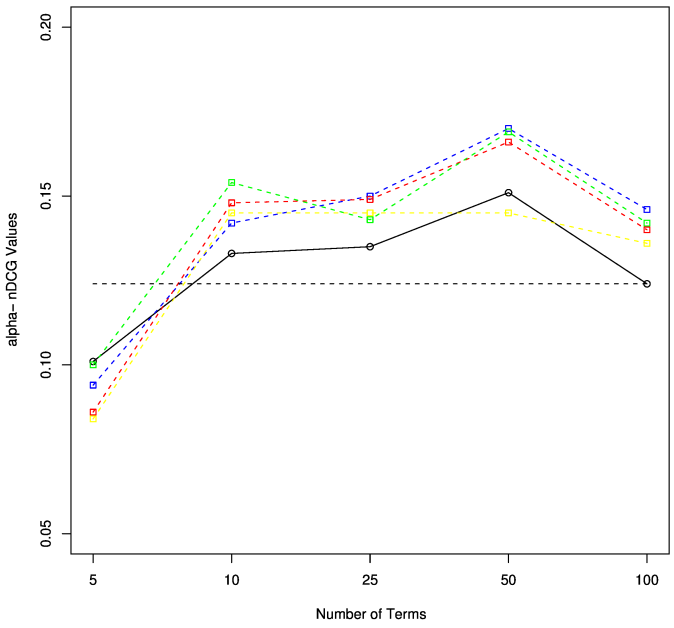

Figure 3.3: α-nDCG averaged over 50 queries with increasing numbers of eigenvectors (subtopic models) and terms in each model.

To investigate the effect of parameters such as m (the number of eigenvectors) and n (number of terms) may affect the performance, we compare the results for a range of values. By comparing the results of the two parameters in Figure 3.3, we see that in general the performance increases and reaches a maximum at 50 eigenvectors and starts to decrease again. The number of terms in the model has less effect on the results.

3.3 Summary

In this chapter, we defined a novel probabilistic model that retrieves a set of subtopics in a small set of documents to be presented to the user. We demonstrated that that our model is competitive with MMR on an intrinsic diversity task, and outperformed another probabilistic model – and all models outperform the traditional IR baselines when tested using a text corpus consisting of news document. We also tested

Run	α-nDCG10	IA-P10
Baseline	0.102	0.056
RM-FM	0.126	0.052
RM-Webgraph	0.168	0.060
SimPrune	0.189	0.081
WUME	0.243	0.131
xQuad	0.282	0.132

Table 3.3: Diversity evaluation results for all our runs sorted by α-NDCG at rank 10

our methods against a web dataset for an extrinsic diversity task and found improvements over a non-diversifying baseline, though explicit subtopic mining techniques and other proposed methods from Section 2.2.2 outperform our methods.

Chapter 4

META-EVALUATION OF NOVELTY AND DIVERSITY EVALUATION

The subtopic framework for novelty and diversity introduced in Section 2.2.2 estimates the degree of novelty in a ranked list using a list of subtopics contained in each document. The subtopic framework defines the novelty of a document in a ranked list by the number of unseen subtopics present in the document. While it is intuitive to break up relevance into smaller pieces, thereby using subtopics to measure novelty and diversity, it is worth analyzing and evaluating the framework in detail.

This chapter takes the necessary steps to evaluate the subtopic-based framework used to evaluate systems that optimize for novelty and diversity. First, in Section 4.1, we statistically analyze some of the common effectiveness metrics described in Section 2.4.2. We take a novel approach using analysis of variance (ANOVA) to measure the relative effect of factors such as relevance and diversity in a ranked list. Second, in Section 4.2, we perform a user study to evaluate the subtopic framework against real user preferences in order to understand the qualities that influence users' preferences in diversity rankings. We introduce a framework for obtaining conditional user preferences to compare a user's notion of novelty to the subtopic framework's definition of novelty through a controlled study.

4.1 Analysis of Evaluation Measures using ANOVA

Analysis of Variance (ANOVA) is a form of statistical hypothesis testing used for analyzing experimental data in which one or more independent variables are measured under various conditions identified by a dependent variable. ANOVA follows a repeated measure design, where experiments are repeated by varying a one or more

dependent variables, thereby partitioning the observed variance in a variable into various components attributable to different sources. For this study, we use ANOVA to determine the degree to which diversity effectiveness metrics such as α-nDCG, ERR-IA, and MAP-IA are influenced by raw levels of relevance and diversity. Our goal here is to decompose the variance in an evaluation measure into the following components:

1. Variance due to changes in the system's ability to find relevant documents
2. Variance due to changes in the ability of a system to satisfy diverse needs
3. Variance due to changes in system's ability to rank relevant and diverse documents
4. Variance due to interactions among the above
5. Variance due to topics
6. Variance due to other attributes of a system or unknown factors

In each of our experiments, we have at least two independent factors from numbers 1–3 above, as well as one random effect (effect due to topics) for which we have repeated measures on every independent factor. The numbers we report are derived from the ANOVA procedures in the statistical programming environment R [174]; they are meant to provide intuition about how much we can distinguish between systems that are different on one factor when the rest are held constant.

4.1.1 Analysis using Real Systems

In order to observe the levels of relevance and diversity on the current systems, we look at various systems submitted to TREC Web track Diversity task. We categorized these systems into three levels of relevance based on precision at rank 10 (with a document judged relevant to any subtopic considered relevant for precision@10) and three levels of diversity based on subtopic recall (S-recall) at rank 10 [197] (which is the ratio of unique subtopics retrieved in the top 10 to total unique subtopics). With three levels of each factor, there were nine categories in total. Figure 4.1 plots S-recall@10 vs precision@10 to show the breakdown of categories in more detail for various runs submitted to the TREC 2009 Web track Diversity task.

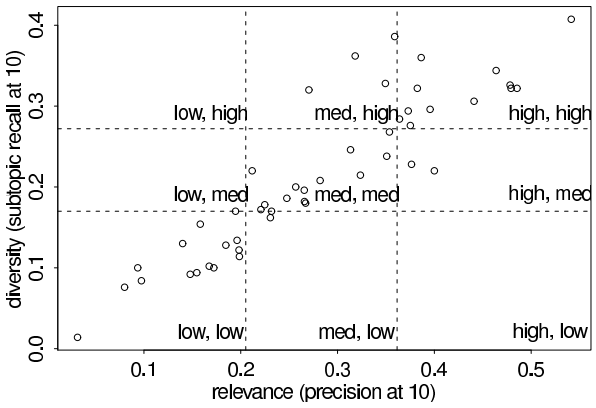

Figure 4.1: Precision@10 vs s-recall@10 for 48 systems submitted to the TREC 09 Web track's diversity task. The dashed lines show relevance and diversity class boundaries.

Note that relevance and diversity among these systems are highly correlated. None of the systems have high relevance and low diversity, nor low relevance and high diversity, though both situations are theoretically possible—high relevance/low diversity could be achieved by a system finding many redundant relevant documents, while low relevance/high diversity could be achieved by a system that finds a few relevant documents covering many subtopics. The fact that few systems fall off the diagonal in the Figure 4.1 suggests that current systems confound relevance and diversity in their ranking approach, and therefore, may not be good for analyzing general properties of measures.

4.1.2 Analysis using Simulated Systems

Since the real systems do not account for all possible scenarios that we want to investigate using our methods, we generated several systems in each category using simulations. We obtained the simulated data by varying independent variables such as relevance, diversity, document ordering, and subtopic distribution. And, the dependent variables in our case are the MAP-IA, α-DCG, and ERR-IA scores. In order to study the effect of independent variable on the evaluation measure, equal number of systems were simulated for each of the nine categories obtained varying the relevance and

diversity levels (as shown in 4.1).

We generated two kinds of simulated systems:

Rel+Div: First, we randomly sampled 10 documents from the full set of relevance judgments to create a random ranking that satisfies one of our nine experimental conditions: low/medium/high precision@10 and low/medium/high S-recall@10, with labels corresponding to values between $0 - 0.3$ for low, $0.3 - 0.6$ for medium, and $0.6 - 1$ for high. There is no special methodology used for this; we simply continue generating simulated rankings until we have generated a minimum of 10 in each of the 9 categories.

Rel+Ord: Next, we controlled diversity ranking in the following way: ten different rankings in each of the same nine relevance/diversity conditions were carefully chosen by varying the minimum rank at which maximum S-recall is obtained. In each category we generate ten rankings in which the documents are re-ordered such that maximum S-recall is obtained only at rank i, where i ranges from 1 to 10. The first ranking (ranking 1, i.e $i = 1$) would attain maximum S-recall at rank 1, the second (ranking 2, i.e $i = 2$) attains max S-recall at rank 2, and so on. In this way, we model degrading ability of a system to rank documents.

4.1.3 Experiment

We perform two sets of experiments. The first studies the influence of relevance and diversity on an evaluation measure. The second experiment tests if the ranking algorithm may play a role in determining the measure value.

4.1.4 Simulation of Systems

To simulate systems, we started with the TREC 2009 dataset (dropping 5 topics that had fewer than three subtopics) and sampled from its 28,000 relevance judgments to generate 10 random rankings for each topic at each pair of relevance and diversity

level by both simulation methods. Thus, we have $3 \cdot 3 \cdot 45 \cdot 10 = 4050$ total data points for our ANOVA, with each of the 45 topics represented in each of the 9 conditions.

4.1.4.1 Varying relevance and diversity

We first study the influence of relevance and diversity on an evaluation measure by varying the precision (relevance) and S-recall (diversity) values at rank 10. In this experiment, we use the system simulated by *Rel+Div* method to obtain 10 simulated rankings for each of the nine categories. Table 4.1 shows ANOVA variance decomposition for our three measures of interest. The first three components (relevance, diversity, and interaction) are independent variables we control. The "topic" component is a random effect due to topic sample, and the "residual" component comprises everything about the measure that cannot be explained by the independent variables. SSE is a measure of the degree to which a component affects variance in the evaluation metric; we have converted SSEs to percentages of total variance such that the column percentages sum to 100%. Larger percentage means that the component has a greater effect on the final value of the metric.

From this table, we conclude the following:

1. α-nDCG does a much better job at distinguishing between systems that provide different levels of diversity, with 52% of its variance being explained by diversity level as compared to 29% for ERR-IA and 20% for MAP-IA respectively.
2. MAP-IA is dominated by random variance due to topic sample. This is because the range of achievable MAP-IAs for a given topic depends heavily on the number and distribution of subtopics in documents [33].
3. ERR-IA is more strongly affected by un-modeled factors captured in residual error than the other two measures. This may imply that ERR-IA is more sensitive to the ranking of documents than α-nDCG or MAP-IA.
4. Interaction between relevance and diversity plays a relatively small role in any of the three measures (though these effects are significant). Our classification of Web track runs suggests interaction effects play a much bigger role in system optimization, however.

Figure 4.2 shows the mean value of each measure increasing with diversity level for each relevance level, with standard error bars showing randomness due to topic

Component	SSE in Evaluation Measure (and %age)		
	ERR-IA	α-nDCG	MAP-IA
Relevance	819.0 (22%)	639.9 (16%)	386.2 (11%)
Diversity	1075.7 (29%)	1979.3 (52%)	648.8 (20%)
Interaction	48.7 (1%)	75.6 (2%)	19.6 (1%)
Topic	482.7 (13%)	567.5 (15%)	1362.8 (42%)
Residual	1282.5 (35%)	561.5 (15%)	822.1 (25%)

Table 4.1: Variance decomposition for components affecting the value of each measure. The first three are independent variables we control. The "topic" component is a random effect due to topic sample. The "residual" component comprises everything about the measure that cannot be explained by the independent variables. Percentages sum to 100 (modulo rounding error) for each measure. All effects are significant with $p < 0.01$.

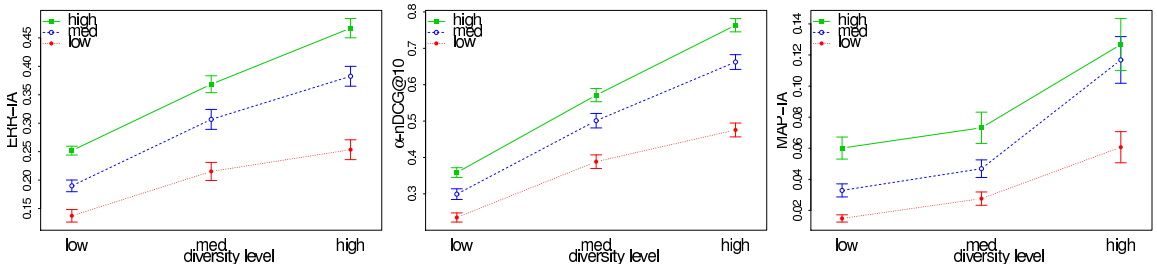

Figure 4.2: Effect of increasing diversity and relevance independently on ERR-IA, α-nDCG, and MAP-IA and their standard error over a topic sample.

sample. This shows that each measure can distinguish between both different levels of relevance and diversity (as ANOVA analysis suggests). Interestingly, standard error tends to increase with diversity and relevance; this suggests that other factors are affecting the measures more when the systems are better.

4.1.4.2 Varying relevance, diversity, and ranking algorithm

In the previous experiment, it was observed the residual error was quite high. Considering that the evaluation measures under study use information about ranks, it is possible that the ranking algorithm may play a role in determining the value returned by the measure. To investigate this effect, we simulated data for different ranking

Component	SSE in Evaluation Measure (and %age)		
	ERR-IA	α-nDCG	MAP-IA
Relevance	682.3 (16%)	586.0 (16%)	386.2 (9%)
Diversity	891.7 (22%)	1174.6 (47%)	648.8 (14%)
Ranking alg	1174.6 (29%)	593.7 (16%)	19.6 (3%)
Interactions	477.5 (12%)	298.3 (8%)	152.9 (3%)
Topic	347.9 (9%)	497.2 (13%)	1362.8 (35%)
Residual	375.0 (12%)	288.1 (7%)	822.1 (35%)

Table 4.2: Variance decomposition for components affecting the value of each measure. The first four are the independent variables we control (interactions between the first three are aggregated together). The "topic" component is a random effect due to topic sample. The "residual" component comprises everything about the measure that cannot be explained by the independent variables or the random effect. Percentages sum to 100 (modulo rounding error) for each measure. All effects are significant with $p < 0.01$.

algorithms; our 10 random rankings above are now non-random levels of a "ranking" factor using the Rel+Ord simulation approach described above in Section 4.1.2.

Table 4.2 summarizes the ANOVA analysis. The first four components are the independent variables we control (interactions between the first three are aggregated together). The "topic" component is a random effect due to topic sample and the "residual" component comprises everything about the measure that cannot be explained by the independent variables or the random effect. We see the same trends as before (Section 4.1.4.1 regarding diversity, relevance, and topic effects, but now we see ranking accounts for a large amount of variance in the ERR-IA and $\alpha - nDCG$. Random variance has decreased from Table 4.1 , except in MAP-IA; this suggests that MAP-IA is dominated by unknown factors.

Figure 4.3 shows the effect of degrading the simulated ranking algorithm on measure value at different diversity levels (averaged over all relevance levels). Note that the maximum ERR-IA values here are much higher than those in Figure 4.2; this is because the ranking of documents is much more important than relevance or diversity alone.

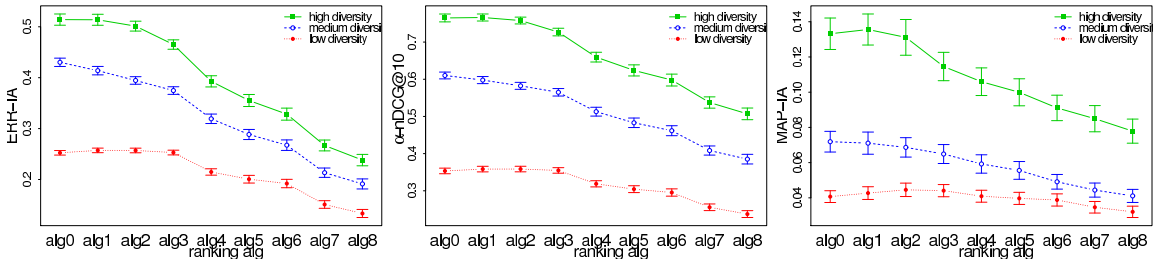

Figure 4.3: Effect of degrading a ranking algorithm at independent diversity levels on ERR-IA, α-nDCG, and MAP-IA and their standard error over a topic sample.

4.1.4.3 Effect of Re-ranking using MMR

The experiment design described above can also be used to investigate commonly used re-ranking approaches for novelty and diversity. As discussed in Section 2.2.1, a common way to achieve diversity in a ranking is to first rank by relevance, then re-rank those documents to achieve greater diversity. We compare two re-ranking approaches: *Maximal Marginal Relevance* (MMR) described in Section 2.2.1 and *Similarity Pruning* (SimPrune) described in Section 3.2.1.3. MMR linearly combines a typical bag-of-words relevance score of a document with the amount of "novelty" the document adds to the ranking [30]. SimPrune is a greedy approach that diversifies the result set by iterating through the initial ranking and removing similar documents [38]. The simulated systems can be used as input to the re-ranking algorithms.

We looked at whether the initial level of relevance and diversity affect the efficacy of the reranking-for-diversity approaches we describe above. We re-ranked results for the random systems using the approaches, then looked at the effect of each of our components on variance in the difference in a measure from the initial ranking to the re-ranked results. Figure 4.4 shows that MMR and SimPrune work best when there is high relevance and medium diversity in the initial ranking, and worst when there is already high diversity in the initial ranking, likely because both tend to exclude documents from the original ranking. The wide range in the error bars shows that in general relevance is not a strong factor, only being significant at $p < 0.1$.

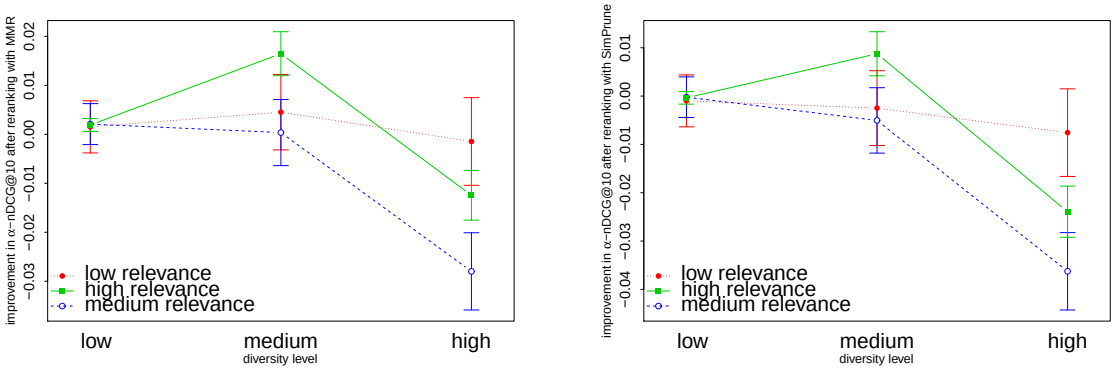

Figure 4.4: Effect on α-nDCG@10 of re-ranking an initial set of results with the given relevance and diversity levels using MMR or SimPrune.

4.1.5 Summary

In summary, we observed that ERR-IA is more sensitive to document ranking and α-nDCG is more sensitive to diversity among the documents retrieved. Also, it is interesting to note that MAP-IA is more sensitive to the topic sample and other factors, which is not desirable in an evaluation measure. The re-ranking approaches were found to be influenced more by diversity in the initial ranking than relevance, with only a medium level of diversity being conducive to improving results after re-ranking.

4.2 Diversity Evaluation vs User Preferences

In the previous section, we analyzed the statistical properties of an evaluation measure for novelty and diversity. Next we conduct a study to measure how properties of these evaluation measures line up with user preferences. Most evaluation measures, including those described in this chapter, require a set of subtopics that disambiguates a given query or provides various pieces of information that comprise the underlying information need. All of these measures estimate effectiveness of a system's ranking by iterating over the ranking, rewarding relevant documents containing unseen subtopic(s) and penalizing relevant documents containing subtopic(s) seen earlier in the ranking. These measures are all based on a few principles in general:

1. A document with more unseen subtopics is worth more than a document with fewer unseen subtopics;

2. A document with both unseen and already-seen subtopics is worth more than a document with only the same unseen subtopic;
3. A document with unseen subtopics is worth more than a document with only redundant subtopics.

We propose a novel method that uses "conditional preferences" to test these principles by conducting a user study. Note that in this study we focus on intrinsic diversity, as it is easier for assessors to understand the concept of relevance when there is less ambiguity of intent.

4.2.1 Conditional User Preferences

A preference judgment is a statement of preference between two documents rather than an absolute judgment of the relevance of a single document [37, 136]. We propose a preference-based framework consisting of a set up in which three relevant documents that we refer to as a *triplet* are displayed such that one of them appears at the top and the other two are displayed as a pair below the top document.

We will use D_T, D_L, and D_R to denote the top, left, and right documents respectively, and a triplet as $\langle D_L, D_R | D_T \rangle$. Figure 4.5 shows an example of such a triplet. An assessor shown such a triplet would be asked to choose which of D_L or D_R they would prefer to see as the *second* document in a ranking given that D_T is first, or in other words, they would express a preference for D_L or D_R conditional on D_T. For the purpose of this study, we assume we have relevance judgments to a topic, and for each relevant document, binary judgments of relevance to a set of subtopics. Thus, we can represent a document as the set of subtopics it has been judged relevant to, e.g. $D_i = \{S_j, S_k\}$ means document i is relevant to subtopics j and k. Varying the number of subtopics in top, left and right documents yields specific hypotheses about preferences for novelty over redundancy.

4.2.2 Hypotheses

The conditional preferences allow us to collect judgments for novelty based on preferences and also enables us to test various hypotheses. Varying the number of

subtopics in D_T, D_L and D_R it is possible to enumerate various hypotheses concerning the effect of subtopics in a document. We define two types of hypotheses: one very specific with respect to subtopic counts, and the other more general.

4.2.2.1 Hypothesis Set 1

First we propose the simplest possible hypotheses that capture the three principles above. We will denote a preference between two documents using $\succ$, e.g. $D_L \succ D_R$ means document D_L is preferred to document D_R. Then the three hypotheses stated formally are:

1. H_1: if $\langle D_L, D_R | D_T \rangle = \langle \{S_2\}, \{S_1\} | \{S_1\} \rangle$, then $D_L \succ D_R$ (i.e., users will prefer a document that contains a new subtopic over a document that contains a redundant one).
2. H_2: if $\langle D_L, D_R | D_T \rangle = \langle \{S_1, S_2\}, \{S_2\} | \{S_1\} \rangle$, then $D_L \succ D_R$ (a user will prefer a document that contains one new subtopic and one redundant subtopic over one that contains only a new subtopic).
3. H_3: if $\langle D_L, D_R | D_T \rangle = \langle \{S_2, S_3\}, \{S_2\} | \{S_1\} \rangle$, then $D_L \succ D_R$ (i.e., a user will prefer a document with two novel subtopics over one with just one novel subtopic).

4.2.2.2 Hypothesis Set 2

Here we define a class of hypotheses in which the number of subtopics contained in each document in a triplet is categorized by relative quantity. We identify six variables based on number of subtopics that almost completely describe the novelty and redundancy present in the triplet. The six variables are as follows:

1. Tn - Number of subtopics in D_T
2. NLn - Number of subtopics in D_L not present in D_T
3. NRn - Number of subtopics in D_R not present in D_T
4. Sn - Number of subtopics that are shared between D_L and D_R
5. RLn - Number of subtopics in D_L and present in D_T
6. RRn - Number of subtopics in D_R and present in D_T

Variable	Number of Subtopics	
	Low	High
Tn	1-4	5-9
Sn	0	1-2
NLn	0-2	3-6
NRn	0-2	3-6
RLn	0	1-2
RRn	0	1-2

Table 4.3: Number of subtopics corresponding to the high and low categories for each variables.

The number of subtopics for each of the six variables are categorized as *low* or *high*. The six variables enable us to test the effect of novelty and redundancy w.r.t the number of subtopics in a triplet. The variables NLn and NRn focus on novelty whereas RLn and RRn focuses on redundancy. For instance, by varying NLn and NRn and holding the other variables constant, it is possible to test the effect of the relative quantity of novel subtopics in a document.

4.2.3 Experiment

We used an online labor marketplace, Amazon Mechanical Turk (AMT), to conduct our experiments. On AMT, "workers" complete HITs (Human Interactive Tasks) that have been submitted by "requestors." A HIT is a small piece of work expected to take no more than a minute or so to complete. Designing a user study using AMT involves deciding on the HIT layout, HIT properties, and quality control measures to control noise in the data. A brief description about each element is given below:

4.2.3.1 HIT Design

Designing a HIT was by far the trickiest part of this user study. In this section, we discuss the variables associated with a HIT and the experimental settings used.

HIT Properties — A detailed description is necessary for the HIT in order to be identified by the workers. In general, workers use the AMT's web interface to search

for a task to work on. Requesters set variables such as HIT Title, Keywords that aid workers to search for tasks that more suitable to their interest.

1. Title: A short description of the task to the workers. The title text is indexed, thus HITs could be searched by title. In our study, *"Document Preference"* was used as the title.

2. Description: A detailed explanation about the task. This gives workers more information before they decide to preview a HIT. The workers can not search based on the description text. We used *"Read the document at the top and pick the document from the two documents shown below that gives most new information"* as the description.

3. Keyword: A set of keywords that will help workers search for HITs. The keywords used in our study are *search, news articles, prefer, preference and opinion.*

4. Time allotted: AMT allows the requester to set a time limit within which a worker has to complete an accepted HIT. It is important not to rush workers into finishing their task. We set *three hours* as the limit to complete a HIT.

5. Pay: Workers are paid for each HIT they complete. Pay rate has obvious implications for attracting workers and incentivizing them to do quality work. Higher pay rates are more attractive to genuine workers but they also attract more spammers. Therefore, care must be taken while determining the pay rates. On the other hand, lower pay rates could result in workers abandoning the task, therefore an appropriate amount needs to be picked. We paid *$0.80* for every HIT used in our study.

HIT Layout — The content of the HIT Design Layout is what a worker sees for a HIT. A common template consisting of various elements was used for all the HITs in the experiment and is shown in Figure 4.5. The various elements used in the template include: a set of instructions about the task, the original keyword query, topic description, article texts (with query keywords highlighted), preference options for indicating which of the two documents the assessor prefers, and a comment field allowing them to provide feedback for that HIT. A brief description about each element is given below:

1. Instructions: The worker was provided with a set of instructions and guidelines prior to judging. The guidelines specified that the worker should assume that everything they know about the topic is in the top document and are trying to find a document that would be most useful for learning more about the topic. Some suggestions included in the guidelines were: one has more new information

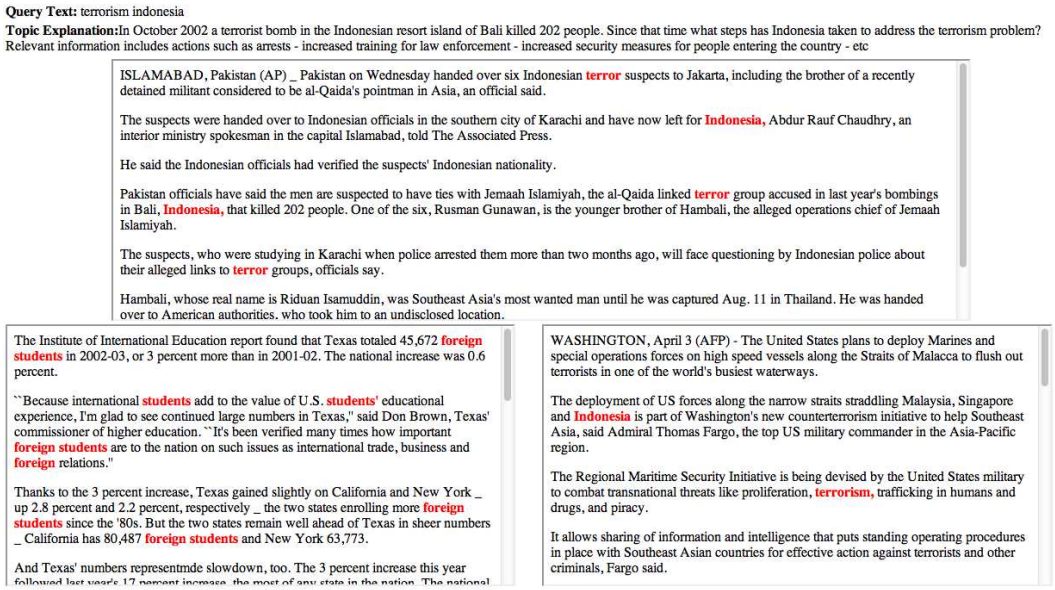

Figure 4.5: Screenshot of the preference triple along with the query text and description.

about the topic than the other; one has more focused new information about the topic than the other; one has more detailed new information than the other; one is easier to read than the other. The actual guidelines used are shown in Figure 4.6.

2. Query text and topic explanation: Each HIT consists of a *query text* field that describes the topic in a few words and a *topic description* field that provides more verbose and informative description about the topic, which are typically expressed in one or two sentences. Below is an example query text and topic description used.

   Query Text: John Kerry endorsement
   Topic Explanation: Documents containing information about individuals/groups that has endorsed or have announced their plan to endorse John Kerry's presidential primary bid are relevant.

3. Preference triplet: Figure 4.5 shows an example preference triplet with the query text and topic description. A HIT consisted of five preference triplets belonging to the same query shown one below the other. Each preference triplet consists of *three* documents, all of which were relevant to the topic. One document appeared at the top; this was a document chosen from the Newswire dataset described in Appendix A, relevant to exactly one subtopic. The bottom two documents in the triplets were chosen randomly such that the hypothesis constraints were satisfied.

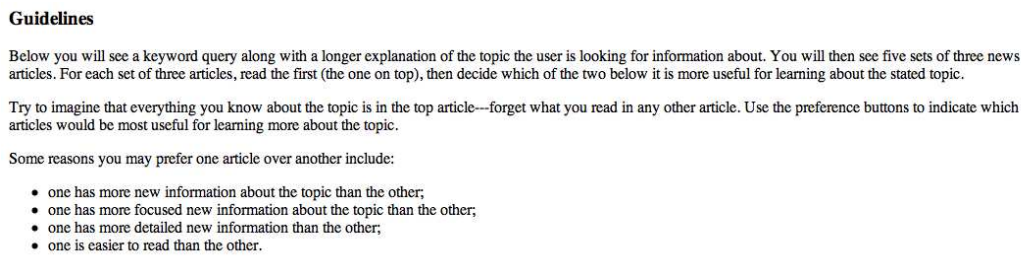

Figure 4.6: Screenshot of the guidelines used in a HIT.

For example, the documents in a triplet for hypothesis *H1* would contain the following subtopics in them: Top Document - S_1, Left Document - S_1, Right Document - S_2.

The workers were asked to pick the document from the lower two that provided the most new information, assuming that all the information they know about the topic is in the top document. They could express a preference based on whatever criteria they liked; we listed some examples in the guidelines. We did not show them any subtopics, nor did we ask them to try to determine subtopics and make a preference based on that. A comment field was provided at the end to provide a common feedback for all the five triplets, if they chose to do so.

Quality Control — There are two major concerns in collecting judgments through crowdsourcing platform such as AMT. One is "Do the workers really understand the task?" and the other is "Are they making a faithful effort to do the work or clicking randomly?". We address these concerns using three techniques: majority vote, trap questions, and qualifications.

Majority vote: Since novelty judgments to be made by the workers are subjective, and it is possible some workers are clicking randomly, having more than one person judge a triplet is common practice to improve the quality of judgments. A variety of methods such as *majority votes* can be used to determine the preferred document in each triplet. In our study, each HIT was judged by 5 different workers and the *majority vote* was used to determine the preferred document.

Trap questions: Triplets for which the answers were already known were included to assess the validity of the results. We included two kinds of trap questions:

"non-relevant document trap" and "identical document trap". For the former, one of the bottom two document was not relevant to the topic. For the latter, the top document and one of the bottom two documents were the same. The assessors were expected to pick the non-identical document as it provides novel and relevant information. One of the five triplets in a HIT was a trap question and the type was chosen randomly.

Qualifications: It is possible to qualify workers before they are allowed to work on your HITs in Amazon Mechanical Turk. Qualifications can be determined based on historical performance of the worker such as percentage of approved HITs. Also, worker's qualification can be based on a short questionnaire. A HIT could have multiple qualifications that a worker must satisfy in order to preview the HIT. A brief description of the two qualifications used in are study is given below:

1. **Approval rate:** HITs can be restricted to workers with a minimum percentage of approval for their task. This method is a commonly used to improving accuracy and reducing spammers from working on your task. A minimum approval rate of *95%* was required by a worker to work on our HITs.

2. **Qualification test:** Qualification tests can be used to ensure that workers have the required skill and knowledge to perform the task. By requiring workers to take a test requester can illustrate the kind of response expected for a task. In our case, workers had to be trained to look for documents that provide novel information given the top document. We created a qualification test having the same design layout as the actual task but had only three triplets. Two of the three triplets were identical document traps and the other was a non-relevant trap. Additionally, we had instructions to the workers for each triplet aiding them in making a preference, e.g. "prefer the document containing information not in the top document" for the identical traps and "prefer the document that is topically relevant" for the the non-relevant traps.

4.2.4 Data

As discussed in Section 4.2, we do not believe these hypotheses would hold for queries with extrinsic diversity (such as those used by the TREC Web tracks), since for those types of queries a user usually has one intent in mind and the rest are not relevant. Thus we need data that models intrinsic diversity, i.e. a user has

H_1	All Prefs		Consensus	
Topic No	Same	New	Same	New
childhood obesity	6	14	1	3
terrorism indonesia	8	12	1	3
earthquakes	15	5	3	1
weapons for urban fighting	15	5	3	1
Total	**44**	**36**	**8**	**8**

Table 4.4: Results for H_1: that novelty is preferred to redundancy. The "all prefs" columns give the number of preferences for the redundant and the novel document for all assessors. The "consensus" columns take a majority vote for each triplet and report the resulting number of preferences.

an unambiguous information need that can be represented by different aspects that appear in relevant documents. An example is "earthquakes" for a user that wants to find locations of recent earthquakes. If there had been earthquakes in Iran, Algeria, India, and Pakistan, and information about them appears in relevant documents, those would be the subtopics.

We have data reflecting this intrinsic diversity need. We refer to the dataset as Newswire data (see Appendix A for a detailed description). It consists of 60 topics, each of which has a keyword query, a description of an information need, and a list of subtopics identified by an assessor. For each topic, 130 documents were judged for topical relevance as well as for relevance to each of the subtopics. The corpus is a set of about 300,000 newswire articles originally part of the AQUAINT corpus.

4.2.4.1 Results and Analysis

Hypothesis Set 1

Judgments for a total of 60 triplets were obtained for hypothesis set 1. The trap triplets used for quality control were not part of the analysis. Since we had each triplet assessed by five separate assessors, a total of 300 judgments were collected out of which 60 were traps.

Table 4.4 shows results for H_1. The "all prefs" columns give the number of

H_2	All Prefs		Consensus	
Topic No	new	same+new	new	same+new
childhood obesity	4	16	0	4
terrorism indonesia	13	7	4	0
kerry endorsement	9	11	2	2
libya sanctions	4	16	0	4
Total	**30**	**50**	**6**	**10**

Table 4.5: Results for H_2: that novelty and redundancy are preferred. The "all prefs" columns give the number of preferences for the redundant+novel document and the novel document for all assessors. The "consensus" columns take a majority vote for each triplet and report the resulting number of preferences.

preferences for the redundant and the novel document for all assessors. The "consensus" columns take a majority vote for each triplet and report the resulting number of preferences. It turns out that there is no clear preference for either redundant or novel documents for the four queries. For two of our queries assessors tended to prefer the novel choice; for the other two they tended to prefer the redundant choice. When we use majority vote to determine a consensus for each triplet, we find that the outcomes are exactly equal. Thus, it is not clear if H_1 holds, we must admit that if it holds it is much less strong than we expected.

Table 4.5 shows a clearer (but still not transparent) preference for H_2, novelty and redundancy together over novelty alone. Over all assessors and all triplets, the preference is significant by a binomial test (50 successes out of 80 trials; $p < 0.05$). Still, there is one query ("john kerry endorsement") for which the difference is insubstantial, and one that has the opposite result ("terrorism indonesia"). The latter case is particularly interesting because it is the opposite of what we would expect after seeing the results in Table 4.4: given that assessors preferred redundant documents to novel documents for that query, why would they prefer novel documents to documents with both novelty and redundancy?

Table 4.6, with results for H_3, is the strongest positive result: a clear preference for documents with two new subtopics over documents with just one. In this case both

H_3	All Prefs		Consensus	
Topic No	new	new+new	new	new+new
childhood obesity	3	17	0	4
terrorism indonesia	2	18	0	4
kerry endorsement	9	11	1	3
libya sanctions	8	12	1	3
Total	**22**	**58**	**2**	**14**

Table 4.6: Results for H_3: that two novel subtopics are preferred to one. The "all prefs" columns give the number of preferences for the novel+novel document and the novel document for all assessors. The "consensus" columns take a majority vote for each triplet and report the resulting number of preferences.

results are significant (58 successes out of 80 trials and $p < 0.0001$ over all triplets and all assessors; 14 successes out of 16 trials and $p < 0.01$ for majority voting). Nevertheless, there are still queries for which the preference is weak.

The results of this experiment show that the presence of novel subtopics *does* influence user preferences, even when they are not explicitly told to look for anything resembling a subtopic. Thus, measures that model subtopics are correctly capturing something that users care about. On the other hand, the degree to which they influence preferences is relatively weak, from statistically insignificant for H_1 to about 75% of the time for H_3. This suggests there are many factors that are important to users, but that these measures are failing to capture.

Hypothesis Set 2

There were a total of 640 triplets (out of which 128 triplets were traps) for the second part of our study. Three separate assessors judged each of these triplets. Thus, a total of 1920 judgments were made out of which 384 were traps. And for this study we had 38 unique workers (identified by worker ID) on AMT working on our triplets. Some of these workers had worked on the first study as well. Almost 70% of the judgments were completed by 15% of the workers and they passed about 93% of the non-relevant traps. The power law distribution for our task has been observed

Topic	High≻Low	Left≻ Right
earthquakes	76 - 20 (79%)	96 - 96 (50%)
terry nichols guilt evidence	75 - 21 (78%)	100 - 92 (52%)
medicare drug coverage	73 - 23 (76%)	86 - 106 (45%)
oil producing countries	65 - 31 (68%)	89 - 103 (46%)
no child left behind	62 - 34 (65%)	81 - 111 (42%)
european union member	61 - 35 (64%)	103 - 89 (54%)
german headscarf court	59 - 37 (61%)	84 - 108 (44%)
ohio highway shooting	51 - 45 (53%)	104 - 88 (54%)
Total	**522 - 246 (68%)**	**743 - 793 (48%)**

Table 4.7: Results of preference judgments by the number of new subtopics in D_L, D_R over D_T (variables NLn, NRn). Counts are aggregated over all values of Tn, Sn per query. The first column gives preference counts for the document with more new subtopics over the document with fewer when $NLn \succ NRn$. The second column is the baseline, giving counts for preferences for left over right.

earlier for other tasks as well [8], we hope to investigate on this issue in the future.

Triplets were generated by controlling four variable: Tn, Sn, NLn and NRn. We obtained sixteen unique settings for the four variable combinations as each of the four variables were categorized into *low* and *high* with equal number of triplets in each setting. This allowed us to perform ANOVA analysis. The number of new subtopics in the left or right document was the primary predictor of preference, with the number of subtopics in the four variables as the secondary predictors. ANOVA indicated that there is a lot of residual variance, suggesting there are various factors influencing preferences that we have not included in the model.

Table 4.7 analyzes preferences for more new subtopics in D_L or D_R over fewer new subtopics (variables NLn and NRn) by topic. We looked at four cases: the first two (NLn high, NRn low; NLn low, NRn high) can tell us whether users prefer to see more new subtopics over fewer, while the second (NLn high, NRn high; NLn low, NLn low) along with the first two give us a baseline preference for left over right. While we would expect the baseline preference to be 50% (since which document appears on the left versus right is randomized), there may be other unmodeled factors that cause

it to be more or less than 50%, so it is useful to compare to this baseline.

It is clear from this table that users as a group prefer to see more new subtopics, just as we saw in the results for H_3 above. Still, there are individual queries for which that preference is not strong, especially when compared to the baseline (e.g. the "Ohio highway shooting" topic), and even when the preference is strong in aggregate there are cases where they do not hold.

There is some effect due to the number of subtopics in D_T, with preferences for more new subtopics stronger when Tn is low. When it is low, the preference for high versus low is 271 to 113 (70%) against a baseline preference for left over right of 347 to 421 (45%)[1]. When Tn is high, the preference for high versus low is 251 to 133 (65%) against a baseline of 396 to 372 (52%).

This second experiment more-or-less confirms the results of the first experiment: that subtopics are an important influence on user preferences, but far from the only salient factor.

4.2.5 Possible Confounding Effects in Display

The way the HITS were displayed may introduce some confounding effects, causing assessors to choose documents for reasons other than novelty or redundancy. In particular:

1. Sometimes the two documents have a large difference in lengths. Assessors may prefer the shorter just to avoid having to read more.

2. Assessors may prefer the document in which more query terms have been highlighted.

3. Assessors may even subconsciously normalize highlighted terms for document length and weight by document frequency, which we could check by looking at preferences due to some retrieval scoring function like language modeling.

We investigated each of these.

[1] We presume that the greater-than-expected preference for the right document is just due to random chance.

4.2.5.1 Document length

It seems that assessors did prefer shorter documents in general, though the preference gets weaker over the three hypotheses. For H_1, assessors preferred the shorter document in 79% of triplets. For H_2, that decreased to 71% of triplets, and for H_3 it dropped steeply to only 44% of triplets. However, it is also true that the mean difference in length for the pair of documents they were choosing between was greatest for H_1 triplets and least for H_3 triplets (158 terms for H_1, 126 terms for H_2, and 47 terms for H_3). It therefore seems safe to conclude that assessors really do prefer shorter documents.

4.2.5.2 Highlighted terms

It turns out that assessors tended to prefer the document with *fewer* highlighted query terms. For H_1, assessors preferred the document with more query terms only 35% of the time. For H_2 that drops to 13%, and for H_3 it comes back up to 29%. The mean difference in number of query term occurrences is quite low, only on the order of one additional occurrence on average for H_1 and H_3 documents, and only 0.2 additional occurrences for H_2 documents. While the effect is significant, it seems unlikely that assessors can pick up on such small differences. We think the effect is more likely due to the distribution of subtopics in documents.

4.2.5.3 Language model score

There was only a slight preference by language model score (using linear smoothing), and it was a preference for documents with a lower score. For H_1, 51% of preferences were for the document with the higher score, but for H_2 and H_3 the preference was 44% and 41% respectively. Since these are not significant, it is unlikely that any interaction between length and query term occurrence had an effect on preferences.

4.2.6 Threats to Validity

In this section, we discuss the major threats to validity in our user study. One of the major threats arises from an assessor's understanding of the task and topics.

Assessors were required to read the guidelines for the task, query, and a description of information need before expressing preference towards a document. Assessor's understanding of the task and their interpretation of the information need could affect their preferences. We used the topic description provided in the dataset; additionally, we manually reviewed the topic descriptions making minor changes to them for clarity and to remove any reference to subtopics.

The results reported in Section 4.2.4.1 are based on comparison of user preferences against the subtopics obtained from a dataset developed by Allan et al. [6]. We assume that the set of subtopics in the dataset provide a reasonable estimate of all possible subtopics for a given query; we believe this is a reasonable assumption as they were originally produced by trained human assessors who were given instructions to find all possible relevant subtopics. Furthermore, we did not give our workers any indication to look for novelty in the form of subtopics, so the signals we detected can be attributed to their actual preferences.

In our study, five different triplets were displayed sequentially in a single HIT for a given query. The order in which the workers viewed the triplets could possibly affect the worker's preference. The position of a triplet was chosen at random to minimize the bias towards a document.

We hired assessors (workers) using a crowdsourcing platform to collect user preferences for our study. The worker population is expected to represent the choices made by real user population using an IR system, which might not be true in reality.

4.2.7 Summary

The results from the user study suggest that the presence of subtopics does influence user preferences, although it is also clear from the analysis that there are other factors strongly affecting preferences. For instance, the results from H_1 and the weaker preference in H_2 were not what we expected. We investigated this more by looking at a number of triplets ourselves and identifying some new hypotheses about why assessors were making the preferences they were. From looking at triplets for

the "earthquakes" topic, we identified three possible reasons for preferring a document with a redundant subtopic:

1. The document updates or corrects information in the top document;
2. The document significantly expands on the information in the top document;
3. Despite containing a novel subtopic, the document provides little information of value.

This suggests to us that there are other factors that affect user preferences, in particular recency, completeness, and value. It may also suggest that there are implicit subtopics (at finer levels of granularity) that the original assessors did not identify, but that make a difference in preferences.

Based on the results of the study, we have reason to believe that users would generally prefer documents with *more* novel information (as quantified by the presence of subtopics). Thus, the subtopic judgments seem to provide a reasonable model of user preferences. However, the subtopic judgments we have may not accurately reflect all the aspects of the topics that users identify. There are reasons for preferences other than novelty and redundancy; these reasons include granularity of subtopics, recency, completeness, value, and perhaps ease of reading (as modeled by document length).

4.3 Summary and Future Directions

We have performed two different meta-evaluations of the subtopic-based diversity evaluation framework. The first statistically analyzed the degree to which relevance, diversity, and document ranking affects three common measures. The second analyzed the degree to which the assumptions these measures are built on are "true" for actual users.

The first analysis suggests that different diversity metrics are, in fact, measuring very different qualities of a ranked list. While this is a positive point in favor of using multiple evaluation measures, the problem is that these measures are so opaque that it is very difficult to understand what differences between them mean. The second analysis suggests that while presence of subtopics is an important reason for user preferences

among documents, there are other factors that play a role as well. The analysis also suggests that various factors including presence of subtopic can be captured implicitly.

Together, these two analyses suggest the need for an evaluation framework that is more transparent and also accounts for a greater variety of reasons for user preferences while still capturing topical relevance and diversity. This is the goal of the next chapter.

Chapter 5

NOVELTY EVALUATION USING USER PREFERENCES

Findings in the previous chapter suggest that user preferences for documents are based not only on the presence of subtopics but also on several other factors: subtopic importance, readability of the document, recency of the document, etc. We expand on them in this chapter, pointing out various issues with subtopic-based evaluation. In Section 5.1, we highlight the disadvantages of the existing subtopic-based measures discussed in Section 2.4.2 using an example that takes into account the task that initiates the search.

Next, in Section 5.2, we develop an assessment framework that measures novelty by allowing users to express preferences without explicitly requiring a set of subtopics, and validate it with a small user study. In Section 5.3, we develop a set of metrics that measure the total utility of a ranked list given a collection of such preference judgments. Finally, we compare our proposed metrics against existing measures using simulations.

5.1 Problems with Subtopic-Based Measures

The subtopic-based evaluation measures (Section 2.4.2) focus on estimating the effectiveness of a system based on topical and sub-topical relevance. In practice, there may be many other factors such as reading level, presentation, completeness, recency, etc. that influence user preferences for one document over another [46]. Even if we decide that it is acceptable to restrict an evaluation to modeling only subtopics, there are some issues with existing measures based on subtopics:

(a) they require a list of subtopics, but subtopic identification is challenging and tricky as it is not easy to enumerate all possible information needs for a given query,

(b) measures such as α-nDCG often require many parameters to be set before use,

(c) the subtopic-based measures (Section 2.4.2) assume subtopics to be independent of each other but in reality this is not true.

	subtopic	*user A*	*user B*	*user C*
a.	What restrictions are there for checked baggage during air travel?			✓
b.	What are the rules for liquids in carry-on luggage?			✓
c.	Find sites that collect statistics and reports about airports		✓	
d.	Find the AAA's website with air travel tips.	✓		
e.	Find the website at the Transportation Security Administration (TSA) that offers air travel tips.	✓		

Table 5.1: An example topic (*air travel information*) along with its subtopics from the TREC Diversity dataset and three possible user profiles indicating the interests of different users.

In order to understand these issues, let us consider an example query from the TREC Web track: *air travel information*. Table 5.1 shows the subtopics defined for the Web track's diversity task and provides the information needs of three different possible users for the given query (assuming we restrict ourselves to representing the user's information need using only subtopics). We can think of user A as a first-time air traveler looking for information on air travel tips and guidelines, user B as a journalist writing an article on the current quality of air travel and looking for statistics and reports to accomplish the task, and user C as an infrequent traveler looking restrictions and rules for check-in and carry-on luggage. Therefore, user A's needs for the above example query consists of subtopics d and e, user B's of c, and user C's of a and b.

First, given the granularity of these subtopics, it would not be difficult to come up with additional subtopics that are not in the data. Top-ranked results from a major search engine suggest subtopics such as "Are airports currently experiencing a high level of delays and cancellations?", "I am disabled and require special consideration for air travel; help me find tips", and "My children are flying alone, I am looking for tips on

how to help them feel comfortable and safe." Are users with these needs going to be satisfied by a system that optimizes for the limited set provided?

Second, measures like α-nDCG and ERR-IA have a substantial number of parameters that must be decided on. Some are explicit, such as α (the penalization for redundancy) [99] or $P(i|q)$ (the probability of an subtopic given a query[1]). Others are implicit, hidden in plain sight because they have "standard" settings: the log discount of α-nDCG or the grade value R_i of ERR-IA, for instance. Each of these parameters requires some value; it is all too easy to fall back on defaults even when they are not appropriate.

Third, some subtopics are clearly more related to each other than others (in fact, we used this similarity to define the users A, B, and C). Documents that are relevant to subtopic c are highly unlikely to also be relevant to any of the other subtopics, but it is more likely that there are pages relevant to both subtopics a and b.

In practice, user satisfaction may be influenced by several factors such as presentation, readability, and other factors as well, but these are ignored by traditional evaluation measures described in Section 2.4.2. The advantage of preference judgments such as those we introduced for our experiment in Section 4.2 is that they allow users to express preferences for any reason that is important to them. And the advantage of our triplet judgments specifically is that they *do* capture some notion of novelty, as our experiments in Section 4.2 show (if not perfectly). Given that they capture to some extent an aspect of system performance that is important to researchers, and also capture aspects of documents that are important to users, we propose to extend this basic idea to a complete framework for evaluation.

5.2 Preference Judgments for Novelty

The idea of pairwise preference judgments is relatively new in the IR literature, having been introduced by Rorvig [136] in 1990, but not subject to empirical study until the past several years [9, 37]. Comparison studies between absolute and preference

[1] The original definition of α-nDCG has parameters for subtopic weights as well.

judgments show that preference judgments capture the notion of relevance to some degree, in that orderings of documents derived from preference judgments correlate to orderings based on absolute judgments. Furthermore, they can often be made faster than graded judgments, with better agreement between assessors (and more consistency with individual assessors) [37]. Using preferences, the assessors can make much finer distinctions between documents. In this section, we discuss in detail a novel preference framework for collecting relevance judgments for the novelty task.

We propose an evaluation framework that simply allows users to express preferences between documents. Their preferences may be based on topical or subtopic relevance, but they may also be based on any other factors that are important to them. Preferences are suitable as they capture varying importance of topics and factors when obtained over many users, and when a sufficiently large set of preferences has been obtained, systems can be evaluated according to how well they satisfy those users. In the following sections we discuss a generalized triplet framework to obtain relevance judgments.

5.2.1 Triplet Framework

Based on previous work showing that preferences correlate to relevance [37] and triplet preferences correlate to novelty (Section 4.2), we now describe a more general triplet framework that involves a series of sets of preference comparisons. Each set of preferences is essentially a comparison sort algorithm, with the comparison function being a simple preference conditional on information contained in top-ranked documents from prior sets of comparisons. Figure 5.3 illustrates conditional preferences with a triplet of documents: the assessor would read document(s) X, and then select one from A or B that they would like to see next. Based on the results in Section 4.2, we expect the assessor's choice to be based not only on topical relevance, but also on the amount of *new* information given what is provided in the top document.

A test collection of preferences for novelty and diversity, then, consists of two different types of preference judgments:

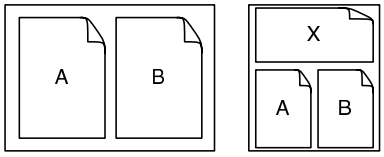

Figure 5.1: Left: a simple pairwise preference for which an assessor chooses A or B. Right: a triplet of documents for conditional preference judgments. An assessor would be asked to choose A or B conditional on having read X.

1. simple pairwise preference judgments, in which a user is shown two documents and asked which they prefer.
2. *conditional* preference judgments, in which a user is shown three or more documents and asked to express a preference between two of them, supposing they had read the others.

Simple pairwise preferences produce an approximation of a relevance ranking: given a choice between two documents, assessors select the one they prefer; as previous work shows, relevance is one aspect that influences their preferences [37]. Since different users may have different needs and different preferences for the same query, pairs can be shown to multiple assessors to get multiple preferences. Over a large space of assessors, we would expect that documents are preferred proportionally according to the relative importance of the subtopics they are relevant to, with various other factors influencing orderings as well.

Simple pairwise preferences cannot capture novelty; in fact, two identical documents should be tied for equal preference in all pairs in which they appear and, therefore, end up tied in the final ordering. Conditional preference judgments attempt to resolve this by asking for a preference for a given pair of document *conditional on* other documents shown to the assessor at the same time. The top document is excluded from this set of preferences.

With this framework, we propose an algorithm for finding the optimal ranking for novelty of a set of documents. First, assessors perform n pairwise preferences. After each preference, the document they selected becomes the one they compare to; in this way the "best" document bubbles up to the top. Assuming preferences are transitive,

only n comparisons are needed to find that "best" document. Once it has been found, assessors perform another set of $n-1$ conditional pairwise preferences, each conditional on the "best" document discovered in the previous round of judging.

The sequence continues in the same way. For the third set, the comparison involves information in *two* previously ranked documents along with a pair of documents; for the fourth, it involves information in three previously ranked documents along with a pair. This continues to the final set, in which there are only two documents to compare conditional on $n-2$ previous top documents. When complete, the most preferred document in the first set takes rank 1, the most preferred document in the second set takes rank 2, and so on. In this way, we obtain the optimal ranking of all documents for a given assessor.

5.2.2 Pilot Study

The triplet framework described in Section 5.2.1 is a general theory. In practice, it could be just as problematic as listing subtopics — in particular, is it feasible to ask for such a large number of judgments? And could assessors tolerate reading m documents at iteration $m-1$ of the algorithm in order to make a preference between two documents they've read many times already? In fact we believe the answer to both questions is "no", but we still wish to see how far we can take the framework.

To that end, we conducted a pilot study with actual users to validate our conditional preference framework. To limit assessor frustration, only two levels of judgments are used (we will resolve the question of how deep we need to go using simulation in Section 5.3.1.6), and the most preferred document in level one is picked as the top-document for level two. Preference judgments are obtained for documents retrieved by the system using an interface described in Section 5.2.2.1 and results are analyzed in Section 5.2.2.2 to demonstrate the feasibility of our approach.

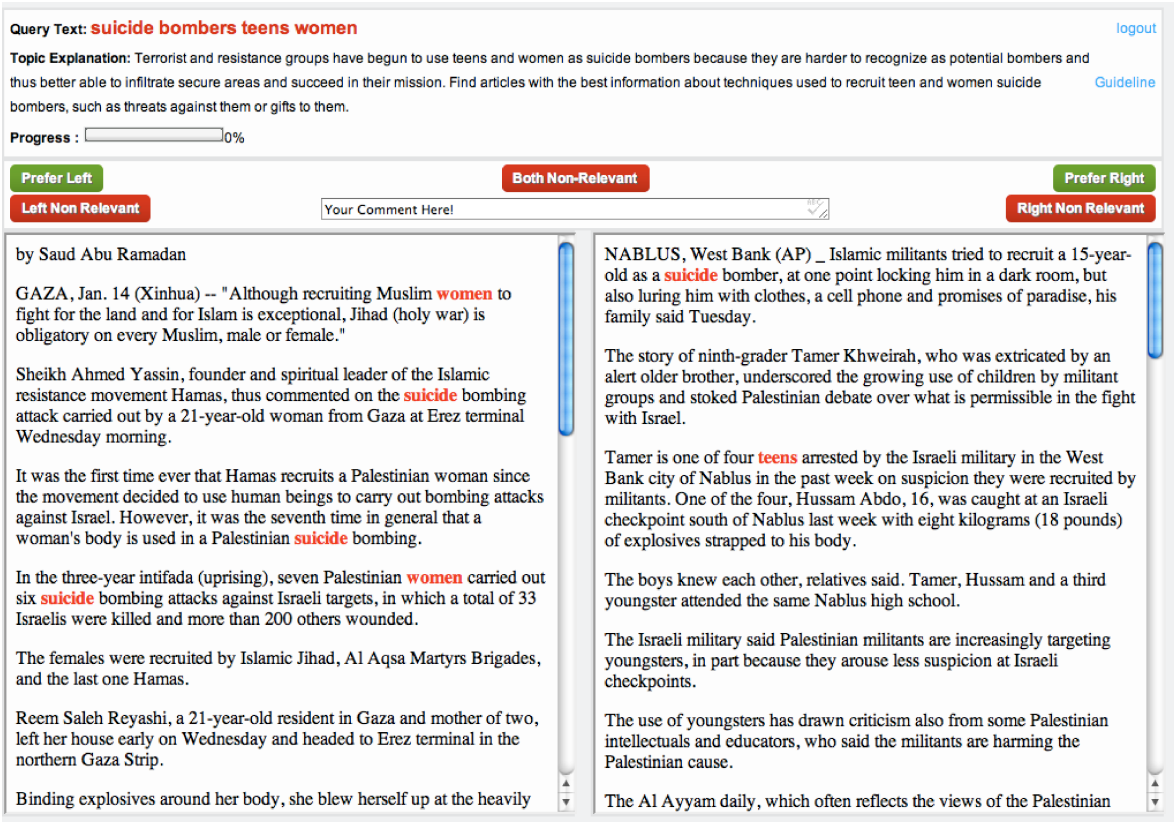

Figure 5.2: Screenshot of the preference collection interface for relevance preferences.

5.2.2.1 Interface Design

We designed a web interface to be used by assessors to collect preferences for both relevance (level 1) and novelty (level 2). Screenshots are shown in Figures 5.2 and 5.3. Common elements in both interfaces are the original keyword query, topic description, article texts (with query keywords highlighted), preference buttons for indicating which of the two documents the assessor prefers, a progress bar with a rough estimate of the percentage of preferences completed, and a comment field allowing them to say why they made their choice (if they wish).

The first two documents shown to an assessor were chosen randomly from the set of all documents to be ranked. After that, whichever document the assessor preferred remained fixed in the interface; only the other document changed. This way the assessor only had to read one new document after each judgment, just as they would in normal

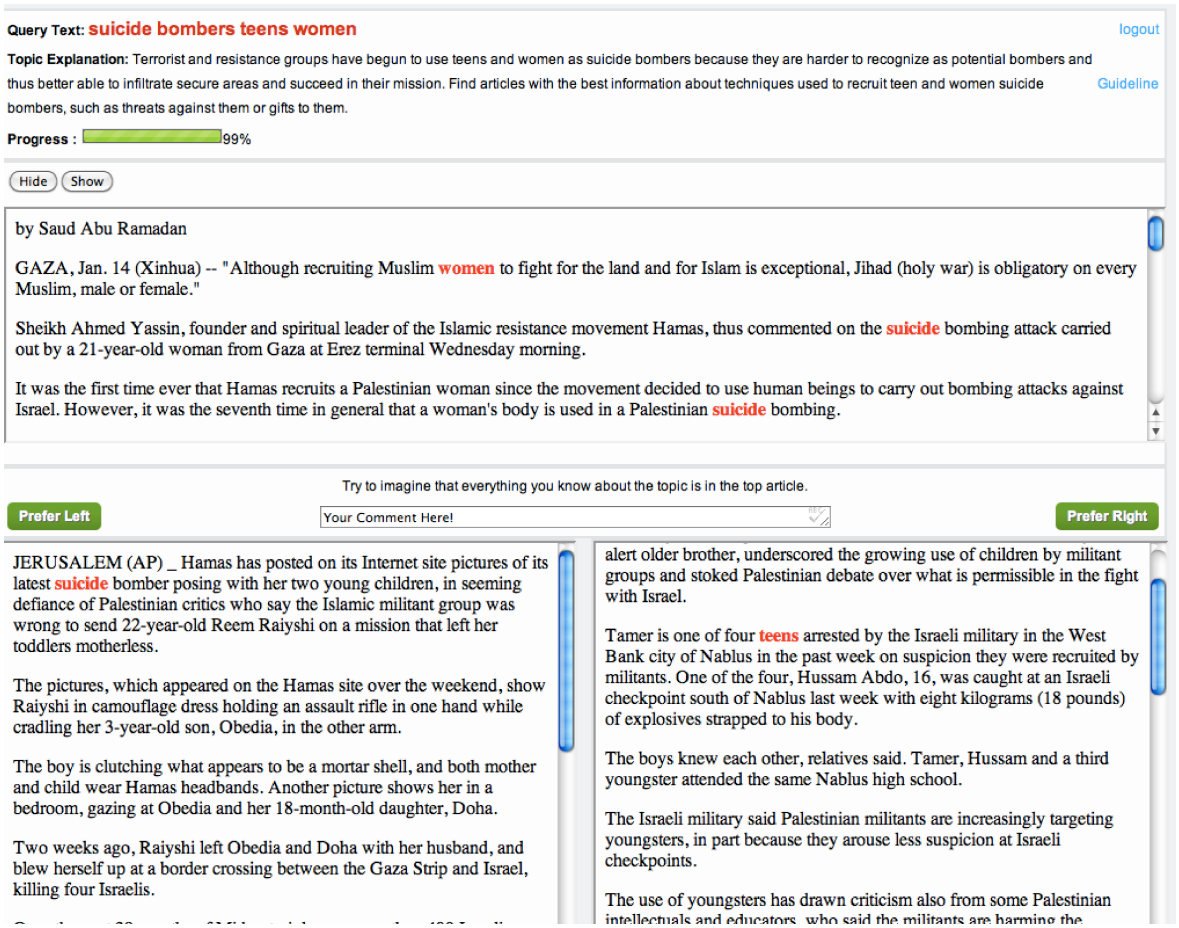

Figure 5.3: Screenshot of the preference interface for the first level of novelty preferences.

single-document assessing setup. Furthermore after the first $O(n)$ judgments we know the top-ranked document for the current set, and thus, if transitivity holds it follows that we only need a linear number of preferences at each set. Since the number of pairwise judgments could be large, assessors could exit to a break at any point and return at the point where they stopped. A progress indicator let them know roughly how close they were to the end (it is a rough estimate because of non-relevant judgments that change the total number of preferences).

First Level Judgments ("Relevance" Preferences): The assessors were shown two documents and a statement of an information need (a topic); the task was to pick

the most preferred document using the "prefer left" or "prefer right" buttons. The assessors were provided with a set of instructions and guidelines prior to judging. The guidelines specified that the assessor should assume they know nothing about the topic and are trying to find documents that are topically relevant; that is, that provide some information about it. If a document contains no topical information, the assessor could judge it "not relevant"; if they do so, the system will assume they prefer every other document to that one and remove it from this set as well as all subsequent sets so it will not be seen in future comparisons. Assessors could also judge "both not relevant" to remove both from the set and see a new pair. If both documents were topically relevant, the assessor could express a preference based on whatever criteria they liked. Some suggestions included in the guidelines were: one document is more focused on the topic than the other; one document has more information about the topic than the other; one document has more detailed information than the other; one document is easier to read than the other.

Second Level Judgments ("Novelty" Preferences): The assessors are shown *three* documents and a statement of an information need (a topic); the task was to pick the most useful document from two of the three to learn more about the topic given what is presented in the third. One document appeared at the top of the screen; this was the most preferred document as identified by the assessor after the first set of preferences. The assessors were asked to pick the document from the other two that provided the most novel information given that they know all the information in the top document. Guidelines specified that the assessor should pretend that the top document is the entirety of what they know about the topic, and their goal is now to find the best document for learning *more* about the topic. Beyond that, they could express a preference based on whatever criteria they liked, including those listed above.

5.2.2.2 Results

We conducted a user study using the above novelty preference interfaces. We used five randomly selected topics from the Newswire data [6] described in Appendix A.

Assessors were computer science graduate students at the University of Delaware. Most were not studying IR, but many were studying NLP and human language technologies. On average, 16.8 documents were judged for each topic. A total of 605 pairs were judged for 4 topics by 8 assessors for experiment levels 1 and 2. We compared these judgments to the original subtopic-based judgments in the data and a brief discussion is given below.

Agreement on Relevance

The performance of our assessors were compared to the original relevance judgments derived from the subtopic judgments. In general a broad agreement on the two classes is 71%. Preference assessors identified 76% of the relevant documents that the original assessors found, and 60% of the documents judged relevant by at least one assessor were judged relevant by both. This is a high level of agreement for information retrieval tasks; compare to the 40% agreement on relevance reported by Voorhees [179].

Transitivity in Preference Judgments

A triplet of documents $\langle i, j, k \rangle$ is transitive if and only if i is preferred to j, j is preferred to k, and i is preferred to k. The ratio of number of triplets found to be transitive to the total number of triplets gives a measure of transitivity in the preference judgments. On average, transitivity holds for 98% of the triplets across all queries with each query being transitive at least 96% of the time. This suggests users are consistent with themselves, and further supports our theoretical model that requires only $O(n)$ comparisons to find the "best" document at any given level.

Rank Correlation

Another way to compare preference judgments to the original subtopic judgments is by using both to construct an optimal ranking of documents, then computing a rank correlation statistic between the two rankings. In order to obtain a ranking of documents from the original subtopic judgments, we simulate preference judgments using the subtopics information. For the first set of comparisons, we always prefer the document with the greatest number of subtopics (in the case of a tie, a random document was preferred). For the second set of comparisons, the most-preferred document from

Topic	Rank Correlation	
	Level 1	Level 2
OPEC actions	0.563	0.534
OPEC actions - Alternate	0.568	0.377
childhood obesity	0.467	0.264
childhood obesity - Alternate	0.403	0.394
suicide bombers teens women	0.320	0.200
foreign students visa restrictions	0.532	0.030

Table 5.2: Kendall's τ correlations between rankings from real preference judgments and rankings from simulated preference judgments (for the relevance ranking (level 1) and the novelty ranking (level 2)).

the first set becomes the "top document", and then for each pair we prefer the document that contains the greatest number of subtopics that are not in that top-ranked document. The final ranking has the most-preferred document from the first set of preferences at rank 1 followed by the ranking obtained from the second set of preferences.

Kendall's τ rank correlation for each topic for both level 1 and level 2 preference judgments is shown in Table 5.2. Kendall's τ ranges from -1 (lists are reversed) to 1 (lists are exactly the same), with 0 indicating essentially a random reordering. Kendall's τ is based on pairwise swaps, and thus can be converted into agreement on pairwise preferences by adding 1 and dividing by 2. When doing this we see that agreement is again high for the relevance ranking, and also high for the novelty ranking, well over the 40% observed by Voorhees. We believe this validates our approach, though certainly the question is not closed.

5.2.3 Threats to Validity

In this section, we briefly discuss the threats to validity in our pilot study. We made use of the interface design discussed in Section 5.2.2.1 to obtain user preferences. The interface needs to be intuitive, and the instructions need to be easily accessible and understandable by novice assessors. A simple and user-friendly interface reduces assessor frustration during the judging process leading to better quality of data. We

started with an interface similar to one used in previous work [37], and relied on feedback from two annotators to create a more intuitive and user friendly interface.

The sample size used in the pilot study was rather small. We used only 4 topics annotated by 6 different assessors, all of them were computer science graduate students. The pilot study was conducted to test the feasibility of our approach and we conduct a large scale user study in Section 6.2 to deal with these issues.

The documents assessors read are news stories from the early 2000s, so the topics may have presented some unfamiliarity to them (though as Table 5.2 suggests, many of these topics are still of interest today). News stories may require some concentration to read and parse, and making a decision about three news stories at once could potentially have a high cognitive load.

5.2.4 Summary

We introduced a general framework for conditional preference judgments and conducted a user study to see if we could find the optimal ranking for a set of documents. The results are positive: users tend to agree with one another about preference; users agree with themselves, meaning transitivity of preferences holds; and the final optimal ranking tends to agree with the one that would be constructed with subtopic judgments. The last point is not as strong, but this is to be expected given that preferences are influenced by factors other than the presence of subtopic as noted in Section 5.1.

Despite some concerns about validity, this experiment is a good sign for our attempts to create an evaluation framework based on conditional preference judgments. It suggests that each assessor can produce a total ordering of documents based on preferences, and moreover that assessors will tend to agree on what that ordering is. This means that automatic systems can be optimized to capture those orderings.

5.3 Preference-Based Measures

The experiment above shows that we can use preference judgments to construct an optimal ranking. If we want to use preference judgments to do more general evaluations of systems, we will need to define new evaluation measures that accept preference judgments rather than absolute subtopic judgments. In this section, we propose a model-based measure using preferences to assess the effectiveness of systems for the novelty and diversity task. As described by Carterette [34], model-based measures can be composed from three underlying models:

- A *browsing model*, which models a user interacting with a ranked list of results. The most accepted model is that a user scans documents down a ranked list one-by-one and stops at some rank k with some probability $P(k)$.
- A model of *document utility*, which tells us how much utility a user derives from a single document.
- A model of *utility accumulation*, which defines the total utility derived from a set of documents seen while browsing the ranked results.

We define our utility based model for novel and diversity ranking task as follows: a user scanning documents down a ranked list derives some utility $U(d)$ from each document and stops at some rank k. We hypothesize that the utility of a document at rank i is dependent on previously ranked document (*i.e.* d_1 to d_{i-1}). Given a probability distribution for a user stoping at rank k, the expected utility can be defined as:

$$Prf = \sum_{k=1}^{n} P(k)U(d_1, ..., d_k) \tag{5.1}$$

where $P(k)$ is the probability that a user stops at rank k and $U(d_1, ..., d_k)$ is the total utility of the documents from ranks 1 through k.

We simplify this by formulating $U(d_1, ..., d_k)$ as a sum of individual document utilities conditional on documents ranked before:

$$Prf = \sum_{k=1}^{n} P(k) \sum_{i=1}^{k} U(d_i|d_1, ..., d_{i-1}) \tag{5.2}$$

where $P(k)$ is the probability that a user stops at rank k, $U(d_i|d_1, ..., d_{i-1})$ gives the utility of the document at rank i conditional on a set of previously ranked documents

from rank $i = 1$ to rank $i = i - 1$, and the sum from $i = 1$ to k gives the total utility of all documents from ranks 1 through k.

There are two main components in the above equation: the probability that a user stops at a given rank ($P(k)$) and the utility of a document conditioned of previously ranked documents ($U(d_i|d_1, ..., d_{i-1})$). Carterette [34] demonstrated different ways to model the stopping rank from the various ad-hoc measure such as Rank Biased Precision [110], nDCG, and Reciprocal Rank.

1. $P_{RBP}(k) = (1 - \theta)^{k-1}\theta$
2. $P_{DCG}(k) = \frac{1}{log(k+1)} - \frac{1}{log(k+2)}$
3. $P_{RR}(k) = \frac{1}{k}$

where k is the rank at which the stopping probability is calculated and θ is a parameter that reflects the patience of users to continue browsing down the ranked list.

Finally, we define the document utility model in which the document utility at a given rank is conditioned on previously ranked documents. The utility of the document at rank i is given by $U(d_i)$ for $i = 1$ since at rank 1 the user would not have seen any other documents and therefore would not be conditioning on any other documents. For subsequent ranks, utility is $U(d_i|d_{i-1}, ...d_1)$, indicating that the utility depends on documents already viewed.

Now our goal is to estimate these utilities using preference judgments. The utility $U(d_i)$ can be directly obtained using the pairwise judgments; we simply compute it as the ratio of number of times a document was preferred to the number of times it appeared in a pair. The utilities $U(d_i|d_{i-1})$ can similarly be obtained from the conditional preferences, computed as the ratio of the number of times d_i was preferred conditional on d_{i-1} appearing as the "given" document to the number of times it appear with d_{i-1} as the "given" document. Note that these utilities can be computed regardless of how many times a document has been seen, how many different assessors have seen it, how much disagreement there is between assessors, and so on. In general, an estimate of a document's utility is obtained using the ratio of the number of times

the document was preferred to the number of times it was shown, in the context of whatever documents were shown above it. An estimate of the document's utility is obtain using the ratio of number of times the document was preferred to the number of time it was shown.

In our pilot study, we only have simple pairwise preferences and conditional preferences in triplets. Thus we cannot directly compute $U(d_i|d_{i-1}, d_{i-2})$ and higher-order dependencies. We need to estimate them somehow from the information we *do* have, namely $U(d_i|d_{i-1}), U(d_i|d_{i-2})$, and $U(d_{i-1}|d_{i-2})$. To do this, we decompose the document utility model as follows:

$$U(d_i|d_1, ..., d_{i-1}) = \begin{cases} U(d_i), & \text{if } i \text{ is } 1 \\ U(d_i|d_{i-1}), & \text{if } i \text{ is } 2 \\ F(\{U(d_i|d_j)\}_{j=1}^{i-1}), & \text{if } i > 2 \end{cases} \tag{5.3}$$

where the function $F()$ takes an array of conditional utilities $(U(d_i|d_j))$.

We experiment with two functions for $F()$: *average* and *minimum*. The intuition behind these functions can be explained with the help of an example. Consider a ranking $R = \{d_1, d_2, d_3\}$. According to equation 5.3 the utility of d_3 depends on $U(d_3|d_1)$ and $U(d_3|d_2)$. The minimum function assumes that d_3 cannot be any more useful conditional on both d_1 *and* d_2 than it is on either one separately, thus giving a sort of worst-case scenario. The average function assumes that the utility of d_3 conditional on both d_1 and d_2 is somewhere in between its utility conditioned on each separately, giving d_3 some benefit of the doubt that it may contribute something more when appearing after both d_1 and d_2 than it does when appearing after either one on its own. These are necessarily approximations. It is possible that d_3 contributes nothing at all to utility after d_1 and d_2, but we have no way of knowing that based on the preferences we have.

Our measure as defined is computed over the entire ranked list. In practice, measures are often computed only to rank 5, 10, or 20 (partially because relevance judgments may not be available deeper than that). When we compute the measure to

a shallower depth, we must normalize it so that it will average over a set of queries. As a final step in the computation of $nPrf$, we normalize equation 5.2 cut off at rank K by the ideal utility score.

$$nPrf[K] = \frac{Prf[K]}{I\text{-}Prf[K]} \tag{5.4}$$

where $I\text{-}Prf[K]$ is the ideal utility score that could be obtained at rank K. This can be obtained by selecting the document with the highest utility value conditioned on previously ranked documents, just as we did in our pilot study. Document (d_1) with the highest utility value takes rank 1 and the document with highest utility when conditioned on d_1 takes rank 2 and so on.

	documents	subtopics a	b	c	d	e
user A	d_1	✓				
	d_2		✓			
user B	d_3			✓		
	d_4			✓		
user C	d_5				✓	
	d_6					✓

List1	List2	
d_1	d_1	
d_2	d_3	
d_3	d_5	
1.0	1.0	α-$nDCG$
0.9	1.0	Preference Measure

Table 5.3: Synthetic example with 6 documents and 5 subtopics. The first ranked list does not satisfy all users where as the second one does but both rankings are scored by equally by α-nDCG, while the preference metrics are able to distinguish the difference.

Table 5.3 provides an example showing the distinction between our preference based measure and α-$nDCG$ based on the user profiles in Table 5.1. The document utilities are estimated by obtaining the preference judgments for all documents from all three users. We would expect the users' preferences to be consistent with their information need, for example user A would prefer d_1 and d_2 consistently to other documents that are not relevant to their needs (but relevant to other needs). Notice that α-$nDCG$ weighs all subtopics equally but the preference measure takes into account the dependency between the subtopics.

5.3.1 Simulation Experiment

In this section, we attempt to validate proposed evaluation measures in Section 5.3 by comparing them to the existing subtopic-based measures. Since those measures do capture both relevance and novelty/diversity, and relevance and novelty/diversity are important to researchers, it is important that our measure capture those aspects as well. Nevertheless, evaluation of the proposed metrics is challenging since there is no ground truth to compare to; there are only other measures. Approaches used in the past to validate newly introduced metrics include comparing the proposed measure to existing measures or click metrics [123, 54]; using user preferences to compare the metrics [153]; and evaluating the metric on various properties such as discriminative power [140]. While each of these approaches have their own advantages, we argue that comparison of existing measures to our measures using simulated data is suitable for evaluating whether our measures capture relevance and novelty/diversity.

Remember, our goal is to build evaluation measures for our preference based framework that assigns utility scores to a document based on user preferences. Generally speaking, when one introduces a new evaluation measure that uses the same types of relevance judgments used in the past, one validates it first by comparing it to existing evaluation measures using the same judgments, then by making additional arguments concerning the need for a new evaluation measure, i.e. what it does differently compared to existing measures. Our case is somewhat different: our measure does not directly admit any type of exiting relevance judgments, but we would still like to be able to validate it against existing measures. Thus, in the following experiments we expect our measures to correlate with the existing subtopic-based measures, since it is important to capture the presence of subtopics. We therefore rely on the existing data with subtopic information to *simulate* user preferences.

5.3.1.1 Data

In our experiments, we used the ClueWeb09 dataset[2]. A total of 150 queries have been developed and judged for the TREC Web track; the number of subtopics for each ranges from 3 to 8. For the diversity task, subtopic level judgments are available for each subtopic indicating the relevance of a document to each subtopic along with the general topical relevance. We also acquired the experimental runs submitted to TREC each year by Web track participants. A total of 48 systems were submitted by 18 groups in 2009, 32 system by 12 groups in 2010, and 62 systems by 16 groups in 2011.

5.3.1.2 Simulation of Users and Preferences

In order to verify and compare our metrics against existing measures, we acquire preferences by simulating them from subtopic relevance information. These will be based on the preferences of simulated users that are modeled by groupings of subtopics (as in Table 5.1). In this way we use only data that is provided as part of the TREC collection, and therefore achieve the fairest and most reproducible possible comparison between evaluation measures.

We simulated user profiles by generating search scenarios for each query and marking subtopics relevant to the scenario. In Section 5.2.1, we explained our reasoning behind the user profiles in Table 5.1 for the query *air travel information*; we use the same approach to obtain the user profiles for all TREC queries. The user profiles were created by the authors and have been made available for public download at `http://ir.cis.udel.edu/~ravichan/data/profiles.tar`. In addition, there is a mega-user that we refer to as the "TREC profile"; this user is equally interested in all subtopics.

These profiles are used to determine the outcome of preferences. For simple pairwise preferences, we always prefer the document with greater number of subtopics

[2] refer to Appendix A for details

relevant to the user profile. In the case of a tie in the number of subtopics, we make a random choice between the left or right document. For conditional preferences, we have three documents (left, right, and top); between the left and the right, we prefer the document that contains the greater number of subtopics relevant to the user profile and not present in the top document. Preference judgments obtained this way are used to compute our preference measure. Finally, using the "TREC profile" to simulate preferences for our measure offers the most direct comparison to other measures.

We have presented a family of preference-based measures for evaluating systems based on novelty and diversity, and outlined the advantages of our metrics over existing subtopic-based measures. In the next section, we demonstrate how our metrics take into account the presence of subtopics implicitly by comparing them with α-nDCG, ERR-IA, and S-recall.

5.3.1.3 System Ranking Comparisons: System Performance

We evaluated all experimental runs submitted to TREC in 2009, 2010, and 2011 using our proposed measure with simulated user profiles as described in the previous section, three different stopping probabilities $P(k)$ and two different utility aggregation functions $F()$. Figure 5.4 shows the performance of systems with respect to both α-nDCG and our preference measure computed with $P_{RBP}(k)$ and $F_{avg}()$ functions and preferences simulated using the "TREC profile". Each point represents a TREC participant system; they are ordered on the x-axis by α-nDCG. Black circles give α-nDCG values as computed by the `ndeval` utility used for the Web track; blue x's indicate the preference measure score for the same system. In these figures we can see that the preference measure is roughly on the same scale as α-nDCG, though typically $0.1 - 0.2$ lower in an absolute sense.

Each increase or drop in the position of x's indicates disagreement with α-nDCG. The increasing trend of the curves in Figure 5.4 indicates that the correlation between the preference measure and α-nDCG is high. A similar trend was observed

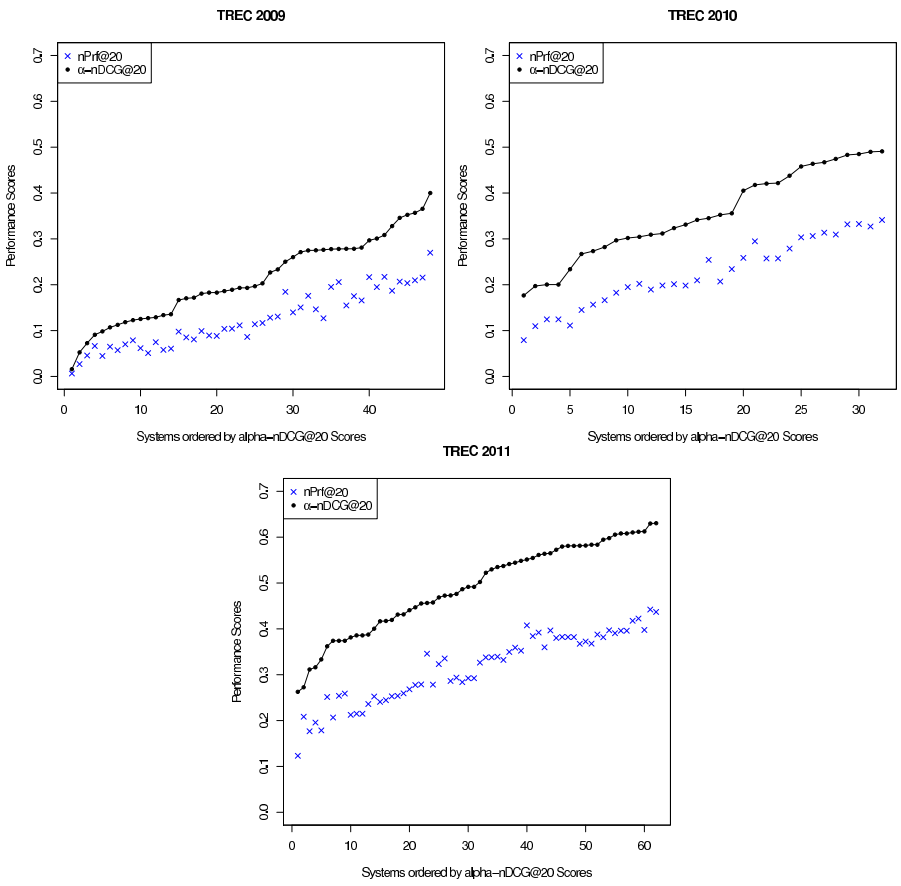

Figure 5.4: TREC 09/10/11 diversity runs evaluated with our preference based metric at rank 20 (nPrf@20) with P_{RBP} and $F_{Average}$. Compare to α-nDCG scores.

while using different $P(k)$ and $F()$ functions as well (not shown). Both α-nDCG and our preference measure agree on the top ranked system in 2009 and 2010.

We analyzed the reason behind disagreement by carefully looking at the actual ranked lists. We investigated how α-nDCG and our proposed measures reward diversified systems on a per topic basis. Based on our analysis, the major reason for disagreement is that α-nDCG penalizes systems that miss documents containing many unique subtopics more harshly than the preference measure does. Much of the variance in α-nDCG scores is due to differences in rank position of the documents with the greatest number of unique subtopics. In practice, this explains the lower scores returned by the preference measure as well.

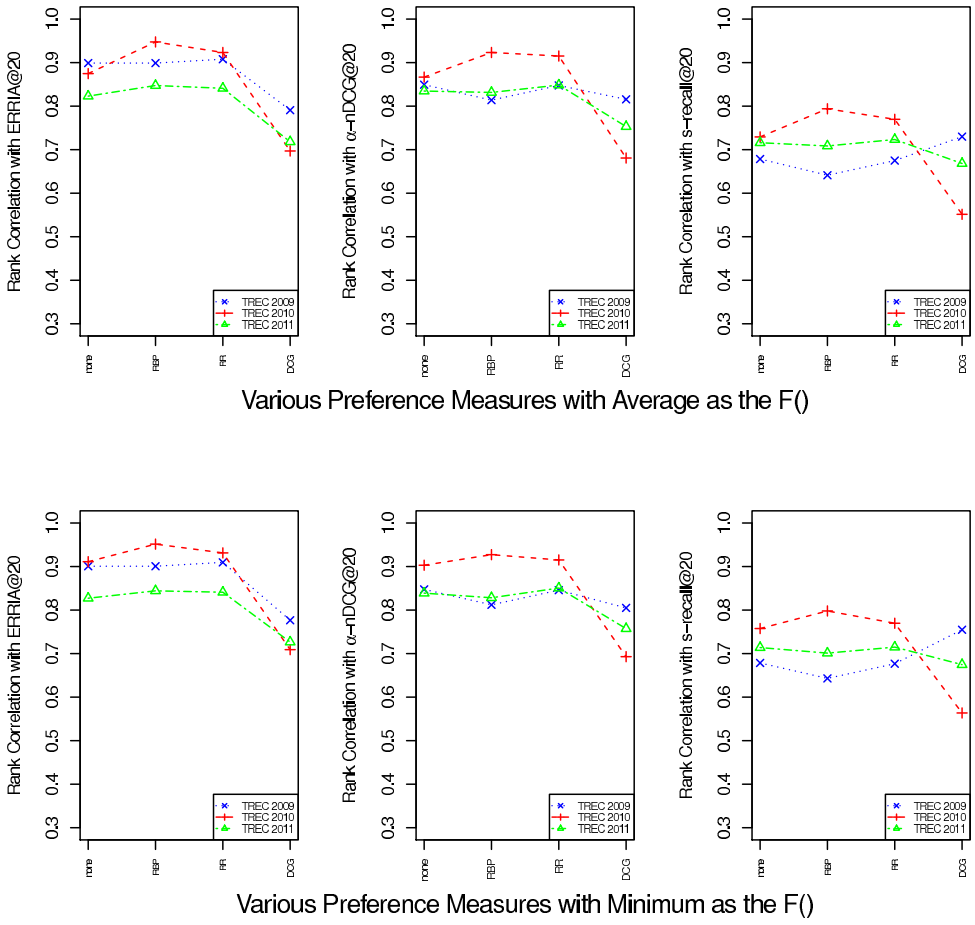

Figure 5.5: Kendall's τ correlation values between our proposed measures and α-nDCG, ERR-IA, S-recall. Values were computed using the submitted runs in the TREC 2009/10/11 dataset. The scores for various $P(k)$ and F() are shown.

5.3.1.4 System Ranking Comparisons: Correlation Between Measures

We measure the stability of our metrics using Kendall's τ by ranking the experimental runs under different effectiveness measures. Kendall's τ ranges from -1 (lists are reversed) to 1 (lists are exactly the same), with 0 indicating essentially a random reordering. Prior work suggest that a τ value of 0.9 or higher between a pair of rankings indicates high similarity between rankings while a value of 0.8 or lower indicates significant difference [24].

Figure 5.5 summarizes the rank correlations between existing subtopic-based

	ERR-IA@20	S-recall@20
α-nDCG@20	0.893	0.828
ERRIA@20	-	0.739

Table 5.4: Kendall's τ correlation values between the existing evaluation measures. Values were computed using 48 submitted runs in TREC 2009 dataset.

metrics and our proposed preference metric using all three $P(k)$ (plus using no $P(k)$ at all—equivalent to a uniform stopping probability) and both $F()$ functions, simulating preferences with the "TREC profile". The correlations are fairly high across TREC datasets, $P(k)$ functions, and $F()$ functions. The $P_{DCG}(k)$ rank function fares worst, with correlations dipping quite a bit for the 2010 data in particular. Subtopic recall is a very simple non-rank based metric for diversity and thus the Kendall's τ values are expected to be slightly lower.

For comparison, Table 5.4 shows the Kendall's τ correlation values between α-nDCG, ERR-IA and S-recall. These correlations are similar to those in Figure 5.5, suggesting that the ranking of systems given by our preference measure varies no more than the rankings of systems given by any two standard measures.

There is almost no difference between the correlations for $F_{avg}()$ and $F_{min}()$ functions for aggregating utility. In fact, the correlation between preference measures computed with those two is nearly 1. Thus we can conclude that the choice of $F()$ (between those two options) does not matter. There is a great deal of difference depending on choice of $P(k)$, however, and thus this is a decision that should be made carefully based on the observed behavior of users.

5.3.1.5 Evaluating Multiple User Profiles

The experiments above are based on the "TREC profile", a user profile that considers every subtopic to be equally relevant. In this experiment, we demonstrate the ability of our methods to handle multiple, more realistic user profiles and show the stability of our metrics. Measures based on absolute subtopic judgments cannot naturally incorporate multiply-judged documents. One must average judgments, or

take a majority vote, or use some other scheme. In contrast, judgments from multiple users can be incorporated easily into our preference framework in the estimation of document utilities, as the document utility is simply the ratio of number of times a document was preferred to the number of times it appeared in a pair, regardless of which user or assessor happened to see it.

We simulate preferences for each of our user profiles for each topic in the TREC set. We compute the preference measure using each profile's preferences separately (giving at least three separate values for each system: one for each user profile), and then use the full set of preferences obtained to compute a single value of the measure. Note that the latter case is *not* the same as computing the preference measure with the "TREC profile": the TREC profile user uses all subtopics to determine the outcome of a preference, while individual users would never use a subtopic that is not relevant to them to determine the outcome of a preference.

We can also compute subtopic-based measures such as α-nDCG against our profiles. To do this, we simply assume that only the subtopics that are relevant to the profile "count" in the measure computation. We will compare values of measures computed this way to our preference measures.

Our hypothesis for this experiment is twofold: 1) that the preference measure computed for a single profile will correlate well to subtopic-based measures computed against the same profile; 2) that the preference measure computed with preferences from all profiles will *not* be the same as an average of the individual profile measures, and also not the same as subtopic-based measures computed as usual. In other words, that the preference measure based on preferences from many different users is measuring something *different* than the preference measure based on preferences from one user, and also different from the subtopic measures.

Figure 5.6 shows the results of evaluating systems using user profile 1, 2, and 3 for each topic and averaging over topics (note that the user profile number is arbitrary; there is nothing connecting user profile 1 for topic 100 to user profile 1 for topic 110). We can see that the system ranking changes for both α-nDCG and the preference

measure, as expected. The correlation between the two remains high: 0.83, 0.88, and 0.82 for user profile 1, 2, and 3 respectively. This is in the same range of correlation values that we saw in Figure 5.5, and supports the first part of our hypothesis.

Figure 5.7 shows the results of evaluating systems with all user profiles, comparing to the evaluation with the TREC profile and with α-nDCG computed with all subtopics. Note here that all three rankings are different, as evidenced by the τ correlations reported in the inset tables. This supports the second part of our hypothesis: that allowing many different users the opportunity to express their preferences can result in a different ranking of systems than treating all assessors as equivalent, as the TREC profile and α-nDCG do.

5.3.1.6 How many levels of judgments are needed?

In Section 5.2.2, we use only two levels of preference for the pilot study. In this section we provide some evidence using preferences simulated from subtopic judgments that two sets of preference are sufficient to approximate an optimal ranking. The relevance ranking is found by preferring a document with more subtopics ("level 1); a first approximation to a novelty ranking ("level 2") is found by preferring a document with the most subtopics that are not in the top document; a second approximation by always preferring a document with the most subtopics that are not in the first two documents ("level 3"); and so on up to level 20.

Figure 5.8 shows the S-recall scores increasing as the number of preference sets increases. Clearly the increase in S-recall from level 1 to level 2 is the largest, nearly exceeding the total increase obtained from all subsequent levels put together. This suggests that the first approximation novelty ranking is likely to be sufficient; this has the benefit of reducing the amount of assessor effort needed to produce the data and also validates our decision to use only two levels in our pilot study.

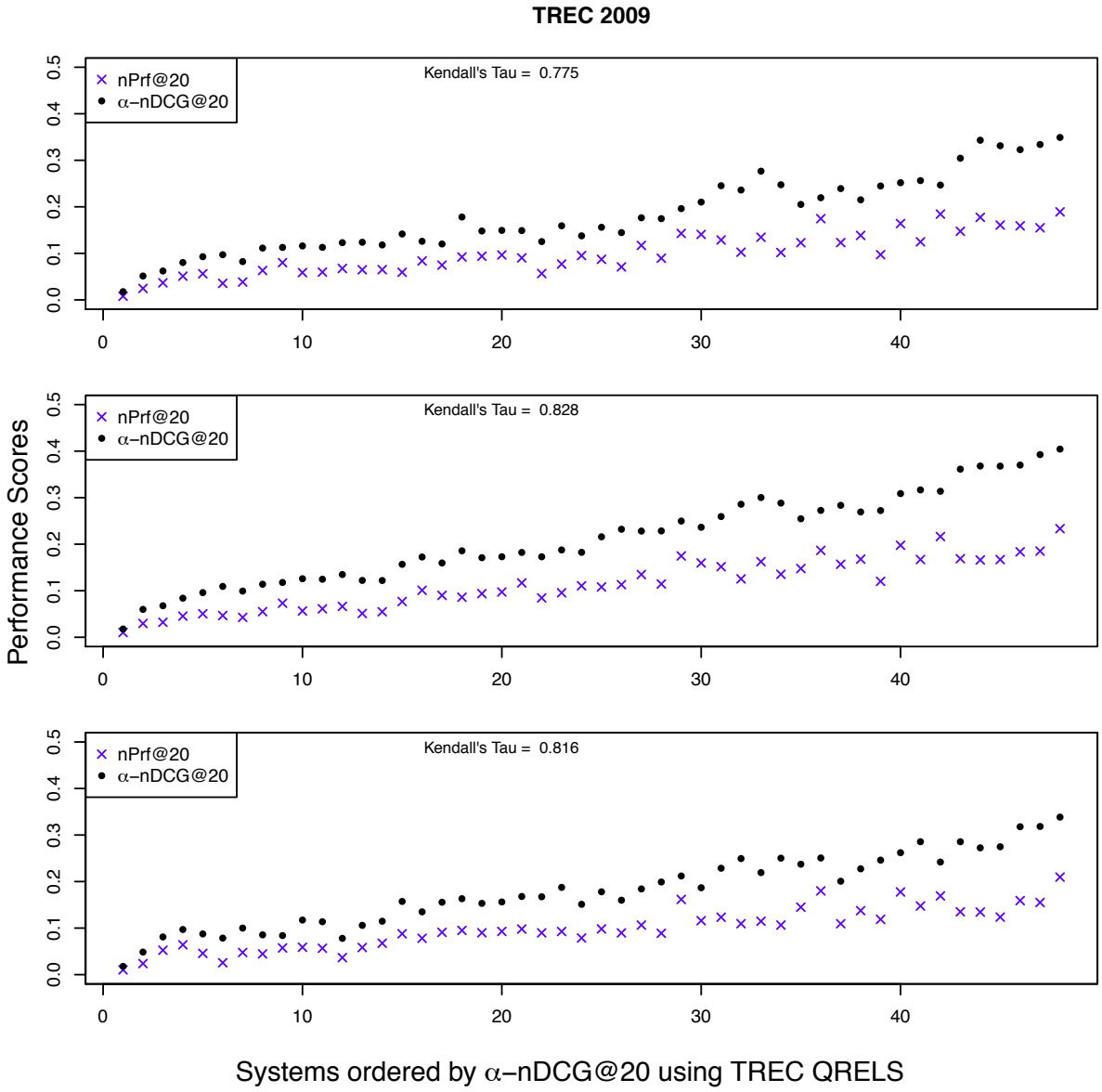

Figure 5.6: Comparison between α-nDCG and our preference measure computed against user profiles 1 (top), 2 (middle), and 3 (bottom) for TREC 2009 systems.

5.4 Summary and Future Directions

In this chapter, we proposed a novel evaluation framework and a family of measures for IR evaluation. The evaluation methodology incorporates novelty and

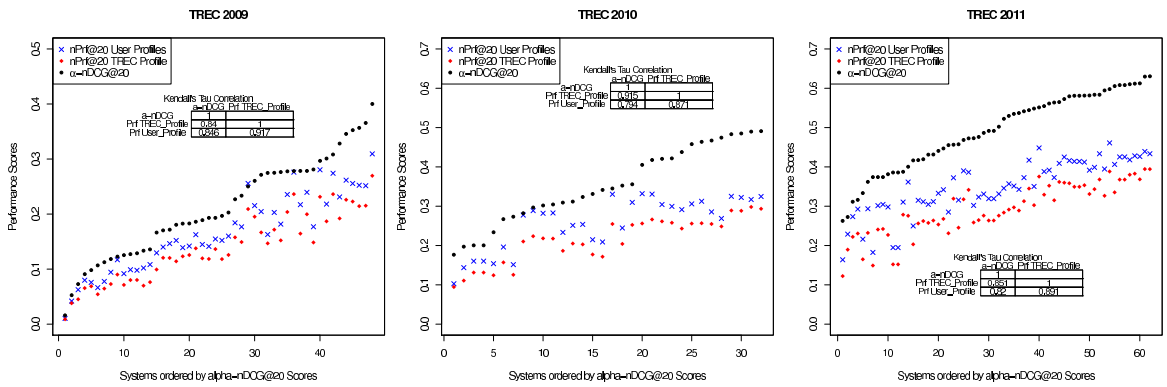

Figure 5.7: Comparison between α-nDCG, our preference measure computed using the TREC profile, and our preference measure computed using a mix of user profiles. Note that all three rankings, while similar, have substantial differences as well.

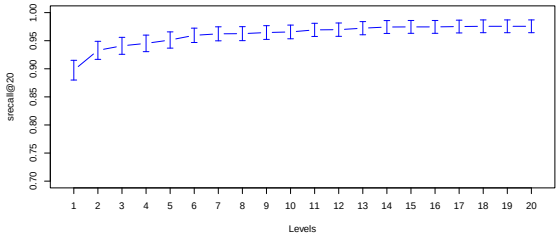

Figure 5.8: S-recall increases as we simulate deeper levels of preference judgments, but the first set of novelty preferences (level 2) gives an increase that nearly exceeds all subsequent levels combined.

diversity, but can also incorporate any property that influences user preferences for one document over another. The measures described are motivated directly by a user model and have several advantage over the existing measures based on explicit subtopic judgments: they capture subtopics implicitly and at finer-grained levels, they account for subtopic importance and dependence as expressed by user preferences, and they require few parameters—only a stopping probability function, for which there are several well-accepted options that can be chosen from by comparing to user log data. They correlate well with existing measures, meaning they do indeed capture something about a system's ability to find relevant and novel material.

Our pilot study shows that it actually is possible for assessors to make preferences in this framework and create usable orderings of documents. Had this experiment failed, the orderings produced would have been entirely unusable, either by not being total orderings or by not being consistent among assessors, or assessors would have quit entirely before completing the task. Our simulation experiment shows that if user preferences are based *entirely* on novelty (as the subtopic judgment framework implies is the most important consideration), then the evaluation measures rank systems in a similar way as existing subtopic-based measures, to as great an extent as any two subtopic-based measures do. Had this experiment failed, correlations between systems would have exhibited a much greater degree of disagreement; that in turn would have suggested that our measures are capturing something very different from the relevance and novelty/diversity aspects that are important to researchers.

TREC style evaluation often involves comparison of a set of retrieval systems, thus requiring relevance judgments for many documents. We believe the use of crowdsourcing platforms are more suitable for our framework, as we would naturally have a large user base with a wide range of preferences. Over a large number of preferences, the most important subtopics and intents would naturally emerge; documents relevant to those would become the documents with the highest utility scores. Yet the conditional judgments would prevent too many documents with those subtopics from reaching the top of the ranking. The measure is designed to handle multiple judgments, disagreements in preferences, *and* novelty of information, and as such it is novel to the information retrieval literature. The clearest direction for future work is to investigate whether our preference measure correlates better with human judgments of system performance than other measures. This is the focus of our next chapter.

Chapter 6

MEASURING SYSTEM EFFECTIVENESS USING USER PREFERENCES

In the previous chapter, we introduced a triplet framework to measure system effectiveness while accounting for inter-document effects, more specifically novelty in a ranked list. The feasibility of our approach for capturing novelty/intrinsic diversity was demonstrated in Section 4.2 and in the pilot study in Section 5.2.2; in the latter we demonstrated that our conditional preferences capture enough about novelty to correlate well with subtopic-based rankings, and in the former we showed that our conditional preferences do indeed capture something about novelty—and that novelty is not the only thing users care about!

The simulation in Section 5.3 showed that we can capture something about extrinsic diversity as well. By grouping similar subtopics together into more cohesive underlying information needs, we capture some diversity in user requests, but far from all of it. We are limited too much by the specific subtopics present in the data we use.

Our aim in this chapter is to unite the threads running through this work: intrinsic vs. extrinsic diversity, user preferences as a way to evaluate systems, and the relative degree to which relevance and diversity are important to a final ranking. To that end, we undertake a large-scale user study, similar to that described in Section 5.2.2 but with many more users expressing more diverse preferences in order to evaluate a much larger set of diversity ranking systems.

6.1 Measuring Effectiveness using User Preferences

Our pilot study in Section 5.2.2 was designed to determine whether users expressing preferences would construct a ranking of documents similar to the "ideal"

ranking one would construct using subtopic judgments; our simulation in Section 5.3.1 was designed to determine whether a preference-based evaluation in which subtopics were the only factor determining preferences would match a subtopic-based evaluation. Our goal in this chapter is to evaluate retrieval systems using actual user preferences. Thus we must make some modifications to the experimental protocol.

The experiment design of the pilot study required both simple pairwise preferences as well as conditional preference judgements. The simple pairwise preferences resulted in the selection of one single "best" document that would then be used for all conditional preferences. Since the pilot study concerned intrinsic diversity, it could be argued that there really is a single "best" document: the one that covers the greatest number of subtopics.

In practice, if extrinsic diversity is in any sense "real", there would be no single document that is better than all the others. Each user would have their own idea of the best document, and it would depend on their own information need, experience, knowledge, and more. Thus it is not necessary to perform the simple pairwise preferences. However, if we are to capture this diversity, we will need more than just one judgment per triplet. We will want to solicit the opinions of a large and diverse user base.

Finally, evaluating a large set of real retrieval systems means we have a much larger pool of potential documents to assess—two systems may be equally "good" according to subtopic-based measures, yet rank very different documents. To be able to distinguish them using our preference measure, we would need to assess triplets consisting of documents retrieved by all systems.

With these changes (triplets with no single "best" document to condition on, multiple user opinions, and many more documents in the assessing pool), we have a cost issue. The number of triplets grows in $\binom{n}{3}$ in the number of documents, and even obtaining as few as 5 preferences per triplet gives a constant multiplier that cannot be ignored: a pool of just 100 documents (which is minuscule in the TREC setting) would require over 800,000 preferences for a complete set, and using the same crowdsourcing

setup we used in the experiment in Section 4.2, would cost over $161,000! (This is part of the reason we relied on simulation for our experiment in Section 5.3.)

Thus putting together a complete set of triplet preferences is out of the question. Instead, we will use only a relatively small set of randomly sampled triplets. Our simple random sampling approach can be described as follows: for each query, a set of documents are pre-selected by pooling the top k documents from the systems we want to evaluate, then all possible triplets from the pre-selected document set are generated. Finally, a subset of N triplets are sampled uniformly at random to be assessed.

An assessor looking at such a triplet could have any of the following cases:

1. All documents in the triplet are relevant to their information need.
2. Only one of the documents in the pair is relevant to their information need (regardless of the relevance of the top document).
3. The top document is non-relevant to their information need.
4. All three documents in the triplet are non-relevant to their information need.
5. Both the documents in the bottom pair are non-relevant to their information need.

For case 1, an assessor can express a preference for any reason they want: one document may be more readable, one may be more recent, one may be more thorough, one may be novel, etc. For case 2, assessors should always pick the relevant document of the two. Case 3 reduces to a preference between the bottom pair of documents; novelty shouldn't enter into it since the top document did not provide any useful information. This is another reason for eliminating the simple pairwise preferences — case 3 subsumes them.

Finally, separate options are provided in the interface to handle cases 4 and 5, where all three documents in the triplet or both documents in the pair are not useful to the assessor's information need.

6.1.1 Validating Random Sampling

Before we began the user study, we used simulations to test our hypothesis of estimating system performance by sampling a small proportion of triplets randomly.

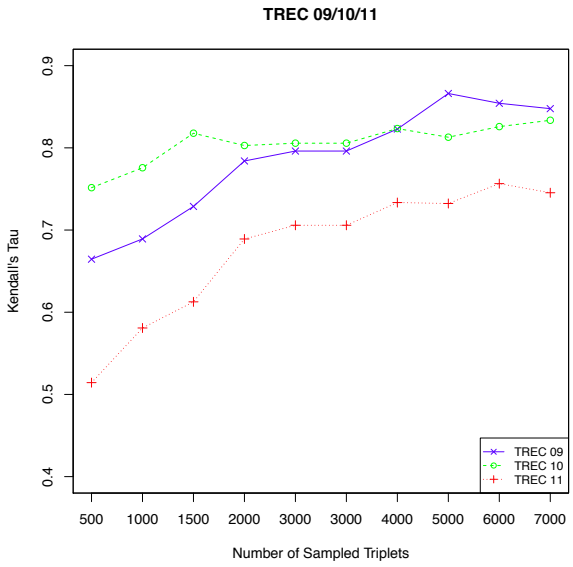

Figure 6.1: TREC 09/10/11 diversity runs evaluated with our preference based metric at rank 20 (nPrf@20) with P_{RR} and $F_{Minimum}$ using single assessor with complete judgments and multiple assessor with incomplete judgments.

We simulate user preference using the same procedure discussed in Section 5.3.1.2, with preferences of simulated users modeled by groupings of subtopics using user profiles (as shown in Table 5.1). For each randomly sampled triplet consisting of three documents (left, right, and top), between the left and the right, the document that contains the greater number of subtopics relevant to the user profile and not present in the top document is preferred.

Our goal is to test the stability of preference measures by comparing the system rankings obtained by using all preferences against a set of incomplete judgments. To do this, we randomly select N triplets of documents for each query. For each triplet, one document is randomly selected to be the "top" document that the other two would be judged conditional on. Though we do not explicitly obtain simple pairwise preferences, we expect that there will be enough cases in which the top document is not relevant to the user profile, and they must fall back on a simple pairwise comparison. We then sample 5 user profiles (with replacement) from those defined for the topic and

simulate their preferences for the triplet. In this way we obtain $5N$ preferences for each topic in a similar way as would be done in a real crowd-sourced assessment. We use those preferences to compute our measure, then compute the correlation to the measure computed with all available preferences. We repeat this 10 times for each topic, measure the correlation each time, and average the correlations.

Figure 6.2 shows the correlation between the system rankings when evaluated using complete judgments and increasing numbers of preferences. Correlation tends to increase as the number of preferences increases, though it does not reach 0.9 (often considered the standard threshold for two rankings to be considered statistically equivalent). This may be partly because simulated user profiles are not evenly represented in the preferences, and partly because our maximum number of preferences is still a fairly small fraction of the total number possible: even selecting triplets from only 100 documents, there are over 161,000 possible triplets, of which we have only obtained less than 5%!

Nevertheless, correlations obtained are consistent with those reported in Table 5.4 and Figure 5.7, which leads us to conclude that it is acceptable to sample a small portion of triplets.

6.2 A Large-Scale Study of Preference Evaluation

To finally bring everything together, we collect a large number of real user preferences using a crowdsourcing platform to demonstrate the ability of our preference metrics to capture both intrinsic and extrinsic diversity in evaluation of automatic systems. As in Section 4.2, we used Amazon Mechanical Turk [1], an online labor marketplace to collect human judgments. The experimental design is similar to the one discussed in Section 4.2.3. The advantage of crowdsourcing is that it provides a diverse user base, which is desirable to capture diverse information needs of a query.

In order to collect data using AMT, triplets had to be organized into HITs (Human Intelligence Tasks). The HIT layout consisted of a set of instructions about the task, original keyword query, topic description, five preference triplets, and a comment

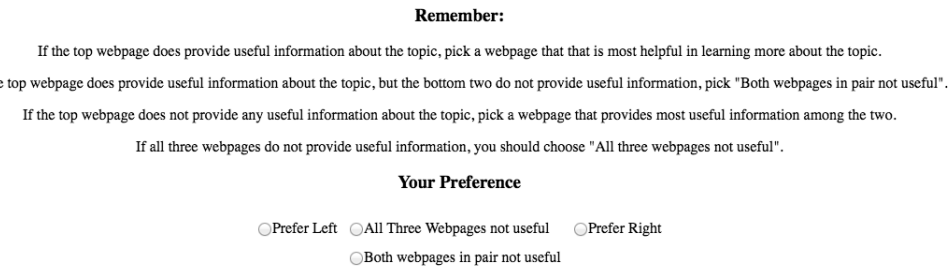

Figure 6.2: Various options available to Amazon Mechanical Turk workers for each of the five triplet in HIT.

field allowing worker to provide feedback. The HITs were identical to the one used in Figure 4.5, although the preference options that followed each triplet to indicate the worker's preference were modified to account for the *five cases* discussed in Section 6.1. Figure 6.2 shows the various options available for each triplet for this experiment. Workers were paid $0.80 for each completed HIT, and each HIT had a time limit of three hours before which it had to be completed.

The quality of data obtained using a crowdsourcing platform such as AMT is always a concern. We included two kinds of "trap" questions to deal with the issue. Traps are simply triplets for which answers should be obvious; each HIT included one trap in order to allow us to assess the validity of the results. The two kinds of trap triplets were "non-relevant document trap" and "identical document trap". For the former, one of the bottom two documents was intentionally selected to be not relevant to the topic (in fact, they were not related to the topic at all); such a document should never be preferred. For the latter, the top document and one of the bottom two documents were exactly identical; the workers were expected to pick the non-identical document as they never see such redundancy in search results. One of the five triplets in every HIT was a trap, but the type was chosen randomly. Workers who failed more than 50% of the trap question were blocked from assessing further HITs to improve the quality of data.

In addition, the HITs were available only to workers who maintained an overall

approval rate of *95%*, and those workers that passed a qualification test (for details of the test refer to Section 4.2.3). This provides additional quality control.

6.2.1 Data

We used the ClueWeb09 dataset[1] that consists of webpages crawled during January and February of 2009. Assessors were shown the same version of the page that is in the ClueWeb09 data. The dataset by default does not contain the supporting files (such as images, CSS stylesheets, JavaScript source, etc) that are necessary for proper rendering of these documents. We had to rely on those still existing on the "live web". In many cases all necessary files were still present at the same location, but some documents rendered poorly, these were removed for this experiment.

A total of 10 queries were randomly selected from the 50 queries used by the TREC 2012 Web track; these are listed in Table 6.1. All 10 queries are rather ambiguous and underspecified. We acquired the ranked results submitted by Web track participants to Web track's Diversity task for the same year. A total of 48 systems were submitted by 12 groups in 2012. We pooled the top 5 documents retrieved by all 48 systems for each of the 10 topics, them sampled 100 triplets from the pool for each topic. User preferences for each of these 100 triplets for each query were obtained from 5 different assessors using the AMT setup described above.

We collected the judgments in batches. For the first batch we obtained judgments for 100 triplets for each of the 10 queries. Table 6.1 gives an overview of this data that was annotated by AMT workers. A total of 5 different workers judged each HIT, this implies that each triplet was judged by 5 different workers. The 100 triplets that were spread across 25 HITs with each HIT containing 1 trap triplet. A total of 125 traps were used for each query (25 HITs $\times$ 5 workers). The number of non-relevant and identical traps failed and average time spent on each HIT are shown in the table. A total of 500 judgments were collected (100 triplet $\times$ 5 annotators) for each query, but only triplets obtained from HIT that passed the traps were used in our experiments.

[1] refer to Appendix A for a detailed description of the dataset

Query Text	Non-Relevant Traps Failed	Identical Traps Failed	Useable Judgments	Pooled Documents
angular cheilitis	5	20	382	39
the beatles rock band	20	22	293	79
septic system design	11	22	368	35
barbados	9	20	382	94
ron howard	14	19	368	54
hip fractures	5	26	361	54
pork tenderloin	8	26	341	55
civil rights movement	5	18	408	73
sore throat	11	30	336	80
fybromyalgia	7	17	384	83

Table 6.1: Overview of the data collected using Amazon Mechanical turk. The documents were pooled (at rank cut-off 5) from systems submitted to TREC 2012 Web Track and 100 triplets were randomly sampled.

The total number of useable judgments for each query is also shown in the Table 6.1. Also shown is the total number of unique documents for each query after pooling, a lower number indicates greater similarity between systems while a higher number reflects systems retrieving documents different from one another. If systems retrieved the exact same documents then the number would be 5 (we used a rank cutoff of 5), whereas if every system retrieved different set of documents then the number would be 240 (48 systems $\times$ 5 documents per system).

6.2.2 Ranking Systems using User Preferences

We evaluated all 48 runs submitted to TREC in 2012 using the data obtained from AMT and our proposed measure with three different stopping probabilities $P(k)$, and two different utility aggregation functions $F()$ (see Section 5.3 for details). Figure 6.3 shows the performance of systems with respect to our preference measure computed with different $P(k)$ functions and two different aggregation functions $F_{avg}()$ (left) and $F_{min}()$ (right). Each point represents a TREC participant system; they are ordered on the x-axis by the preference measure with $P_{DCG}(k)$ and $F_{avg}()$ for the graph

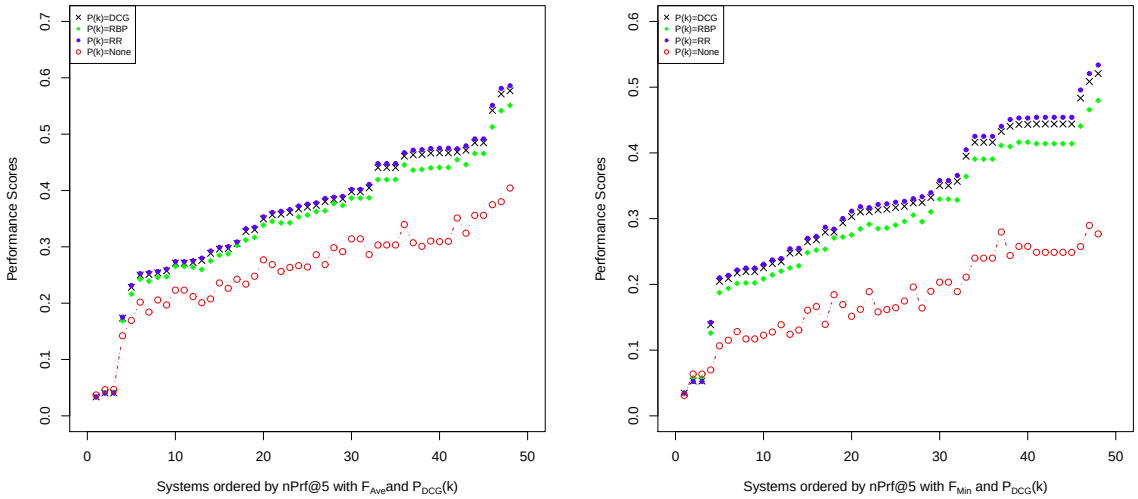

Figure 6.3: TREC 2012 diversity runs evaluated with our preference based metric at rank 5 (nPrf@5) with different $P(k)$ and $F()$.

on the left and $F_{min}()$ the left figure. Though we have five separate assessments for each triplet, we do not need to resolve disagreements; we just provide the measure with all of the raw data and allow it to use that data as it may.

Red circles give $nPrf@5$ values as computed with $P_{None}(k)$; blue xs indicates a preference measure with $P_{RR}(k)$ and so on. Each increase or drop in the position of the circles or dots indicate disagreement with preference measure computed with $F_{ave}()$ and $P_{DCG}(k)$. The increasing trend of the curves in Figure 6.3 indicates that the correlation between different implementations of the preference measure is high. All measures that use $F_{ave}()$ agree on the top ranked system ($uogTrA44xu$), whereas the top ranked system was $QUTparaTQEg1$ when evaluated using $F_{min}()$ aggregation function. The system $QUTparaTQEg1$ was the second ranked system with $F_{ave}()$. In general, high correlation of around $\tau = 0.9$ was observed between measures that used the same discount function $P(k)$ but different aggregation function $F()$, expect for $P_{None}(k)$. The correlation drops to $\tau = 0.79$ between a measure that used $F_{min}()$ and $F_{ave}()$ with no discount function $P_{None}(k)$. While this is low, it is not unexpected.

6.2.3 Comparison with TREC Data

We now compare our approach against the diversity measures that rely on subtopic-level relevance judgments. Since we used the TREC 2012 Web track data (topics and submitted systems), subtopic level judgments were available for each document in every triplet we selected for every query. This enabled us to compare the traditional subtopic-based approach with our triplet-based approach. First, we compare the two approaches in terms of how well our preference judgments "agree with" judgments based on subtopics, and then we compare our preference-based metrics against diversity measures such as $\alpha - nDCG$, S-recall and ERR-IA.

6.2.3.1 Triplet Comparisons

In order to compare the user preferences obtained using the triplet framework from AMT against TREC preferences that are based on subtopics, we rely on simulations similar to those used in previous experiments. Since our preference judgments provide no information about subtopics, we must simulate preference judgments from the TREC subtopic judgments.

We simulate TREC user preferences following a set of principles that are common to subtopic-based diversity measures such as α-nDCG and S-recall. Preference are simulated for each triplet as follows:

- If the top document is relevant – a document that contains more unseen subtopics (i.e. subtopics not in the top document) is preferred to a document with fewer subtopics.
- If the top document is relevant – a document that contains more unseen subtopics (i.e. subtopics not in the top document) *and* redundant subtopics is preferred to a document with only the same unseen subtopics.
- If the top document is relevant – a document that contains more unseen subtopics (i.e. subtopics not in the top document) *and* redundant subtopics is preferred to a document with only redundant subtopics.
- If the top document is non-relevant – a document that contains more subtopics is preferred to a document with fewer.

- Triplets with all three non-relevant documents and triplets with both non-relevant documents in the pair were explicitly indicated as such.

Using these rules, we obtain simulated preferences for all 100 triplets assessed for all 10 queries. To compute agreement, we calculated the percentage of triplets for which the user preference matched the TREC preference. Note that in the simulation, there could be cases where no preference can be made ("ties"), while the Mechanical Turk workers were forced to make a preference in every case. Therefore, any preference made by AMT worker was considered as agreement as long as they agreed with TREC preference on relevance of the documents.

Table 6.2 provides the percentage agreement between the two sets of preferences for each query. The agreement percentages observed are comparable to the results of experiments conducted on the *Newswire dataset* in Section 4.2.4.1. Overall they are quite high, yet low enough that we can assume assessors really do have different preferences.

query	agreement
angular cheilitis	75.1%
the beatles rock band	63.1%
septic system design	56.2%
barbados	57.6%
ron howard	55.7%
hip fractures	58.7%
pork tenderloin	72.7%
civil right movement	63.7%
sore throat	54.5%
fybromyalgia	60.7%
total	**61.8%**

Table 6.2: Agreement percentages between TREC subtopic preferences and user preferences.

6.2.3.2 Comparison to Rankings by TREC Measures

The triplet level comparison shows there is considerable disagreement between TREC preferences and user preferences. The important question is whether these

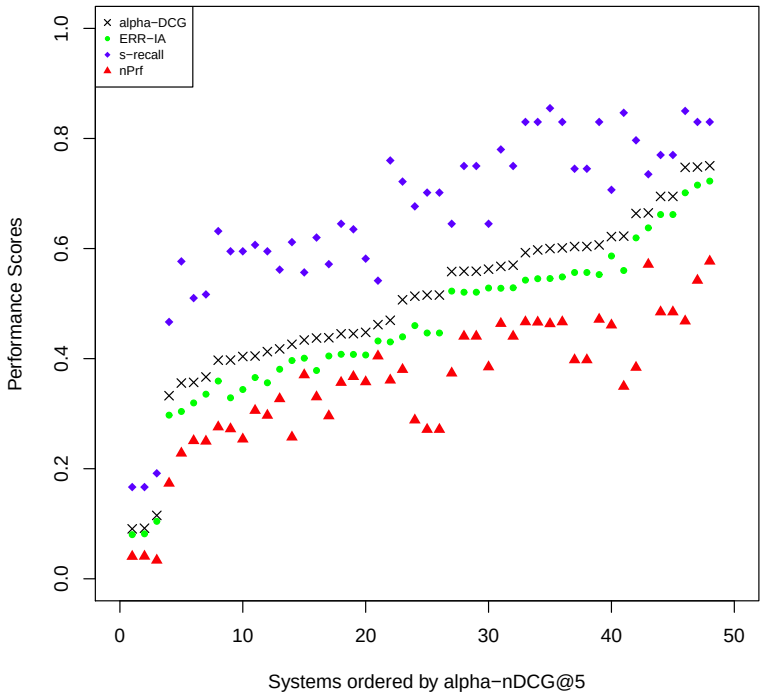

Figure 6.4: Kendall's τ correlation between various TREC measures evaluated using subtopic judgments and preference-based metrics computed using user preferences.

difference affect system ranking. Thus, we evaluated all experimental runs using both the TREC measures as well as our proposed measure with three different stopping probabilities $P(k)$ and two different utility aggregation functions $F()$.

Figure 6.3 shows the performance of systems with respect to both α-nDCG (averaged over the same five topics) and our preference measure computed with $P_{DCG}(k)$ and $F_{avg}()$ function. For comparison, the figure also shows the scores computed using ERR-IA and S-recall at rank 5. Each increase or drop in the position of dots, triangles and diamonds indicate a disagreement with α-nDCG. The top ranking system was the same for α-nDCG, ERR-IA, and nPrf.

Table 6.4 gives the Kendall's τ correlation between existing evaluation measures and our preference evaluation using user preferences. This suggests that the ranking of

	ERR-IA	α-nDCG	S-recall
nPrf	0.73	0.74	0.62
S-recall	0.70	0.71	
α-nDCG	0.96		

Table 6.3: Comparison of system rankings evaluated using different existing evaluation measures using subtopic judgments and preference measures using user preferences. Values were computed using 48 runs submitted to the TREC 2012 Web track.

systems given by our preference measure varies no more than the rankings of systems given by S-recall and other measures. In fact, the correlation is very high considering that:

- only a tiny fraction of all possible triplets was assessed for the preference measure;
- our assessors, unlike TREC assessors, are not trained and were not given precise guidelines for judging;
- our assessors were not given any information at all about the subtopics used for the TREC measures;
- our assessors were free to give preferences for any reason they wanted, not just relevance and novelty;
- the web pages our assessors saw were not identical to those seen by TREC's assessors, and in some cases may have been quite different (due to the problem of support files mentioned above).

Considering how different the two evaluation scenarios are, it is quite remarkable that the system rankings would agree so highly.

6.2.3.3 Analysis

The disagreement with α-nDCG is lower for poorly performing systems and much higher for high performing systems, possibly indicating that our approach treats the relevant documents in a ranked list differently from existing subtopic based approach. The triplet level comparison experiment showed that presence of subtopics in a document is captured indirectly and there are other factors that influence user

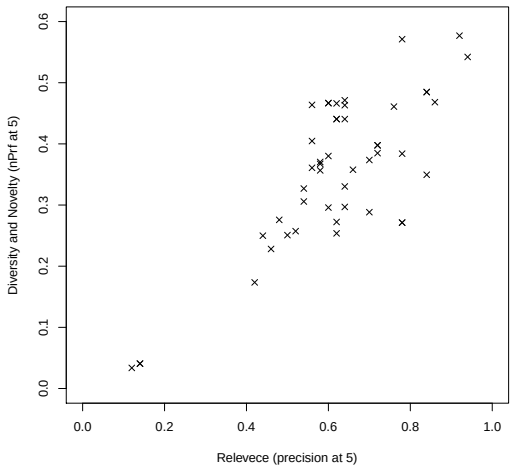

Figure 6.5: Comparison of runs under traditional ad-hoc (measured using Precision@5) and diversity (measured using nPrf@5) effectiveness measures.

preference. These factors could certainly be a cause for disagreement between $nPrf$ and α-nDCG.

An evaluation measure designed to capture novelty and diversity in a ranked list must distinguish between a system that simply retrieves any relevant document and a system that retrieves documents containing novel and relevant information. Figure 6.5 plots nPrf@5 vs precision@10 to show the correlation between our preference measure and relevance. An upward trend can be observed indicating correlation between relevance and diversity, with Kendall's τ correlation of 0.47. The correlation is much higher for systems with low relevance; a Kendall's τ of 0.77 was observed between relevance and diversity when only systems below the median performance ($Prec@5 < 0.62$) were considered. The correlation drops to 0.28 when systems above the median were considered. This is a clear indication that our approach does something different from traditional ad-hoc measures only when relevance is high in the ranked list, which is desirable for a measure that captures novelty and diversity.

6.2.4 Threats to Validity

In order to generalize the findings made in our study, the topics (artificially created information needs) are expected to be representative of actual needs encountered by real users using an IR system. We rely on topics developed for TREC Web track Diversity task, which was constructed by sampling queries from logs of a commercial search engine; these topics are widely used in various works, so even if they are not representative they are a common standard in the literature [55, 53, 56, 59].

The guidelines provided to assessors needs to be clear and easily understandable. We rely on two graduate students from computer science to provide feedback and made changes to the guidelines based on their feedback. Additionally, a comment box was included for workers to provide feedback, although we did not get any questions regarding the task.

The assessor used to obtain user preferences are expected to represent real user population using an IR system. We rely on workers from Amazon Mechanical Turk to obtain our user preferences. The use of crowdsourced workers to obtain relevance judgments is widely used in the IR community [7], although it is not clear how representative they are of a real user population. The reliability of crowdsourced workers providing user preference is always a concern, random choices made could lead to poor or inaccurate data. We use the quality control criteria discussed in Section 6.2 to determine the reliability of a worker. A HIT sequentially displays five different triplet of documents on the same query and users often judge triplets sequentially. The manner in which triplets are displayed could cause a learning effect, thereby inducing bias into the preferences made. The position of the triplet in a HIT is picked randomly to minimize any selection bias.

The most important threat to validity is our assumption that it is good for our measure to be highly correlated to TREC measures. Certainly some correlation is expected; as we have shown in experiments throughout this work, our preference judgments do capture something about both relevance and novelty/diversity and therefore

are expected to correlate to measures that are designed to capture those aspects explicitly. We emphasize that it is not at all clear how high that correlation "should" be. We certainly do not want it to be perfect, as that would imply that the subtopic framework perfectly captures user preferences. We also do not want it to be excessively low—Kendall's τ correlations lower than 0.5 would be far too low to conclude that our measure is reliably capturing relevance and novelty/diversity, since that is about the minimum expected when relevance is just assigned to documents randomly [167]. Thus it seems to us that the correlations we observe, which sit between those two undesirable extremes and are also concordant with correlations between existing measures, can be qualifiedly referred to as good.

6.3 Summary

In this chapter we undertook a relatively large-scale user study in order to determine whether our approach could be feasibly used in an evaluation environment like TREC. Despite the large number of triplets required for a complete set of assessments, it seems the answer is yes: our measure agrees with subtopic-based evaluation measures on the best-performing systems (which is important for researchers trying to optimize a system); moreover, it is highly self-consistent even when its parameters are changed (as shown in Section 6.1.1). It correlates as well with TREC measures as S-recall does, highly enough that we can conclude it is capturing something about both relevance and intrinsic/extrinsic diversity, but not so high that it is measuring measuring nothing but the same qualities that existing measure are capturing.

Furthermore, our approach does not suffer from the problems those measures do. We require no enumeration of subtopics; we only need to provide assessors with a query and let them express their preferences according to whatever they believe the intent of the query is. With multiple assessors giving judgments on the same preferences, dominant intents will rise; since our measure admits multiple assessments without having to resolve disagreements, it naturally captures this diversity. While our measure has parameters that must be set, it is consistent with itself across many

reasonable settings of those parameters. Finally, our measure makes no assumptions about independence of subtopics or anything else about why one document should be preferred to another; users are free to give their preference for any reason that is important to them.

Chapter 7

CONCLUSION

We began by introducing a novel probabilistic set-based framework to select a small set of relevant documents that covers various aspects of a given query. We proposed three different to hypothesize subtopics for a given query: building language models using pseudo-relevant documents, webgraphs and topic models. Evaluation of this model on two different data sets suggests that there is a difference in how retrieval models should optimize for intrinsic versus extrinsic diversity, which in turn suggests that different evaluations might be needed—or a unified approach that implicitly captures both.

To that end, we have presented a novel information retrieval evaluation framework based on conditional pairwise preferences. The framework is designed to indirectly capture the amount of relevance, novelty, and diversity contributed by a document to the ranked list, as well as any other factors that are important to users. We proposed a set of evaluation metrics to estimate the total utility of a ranked list using an aggregation function and a discount function. There are several advantages of this proposed approach: it is motivated directly by a user model, it captures subtopics implicitly and at finer-grained levels than is possible with explicitly-listed subtopics, it accounts for subtopic importance and dependence, it can directly handle disagreeing judgments from multiple assessors, and it requires few parameters – only a stopping probability function.

We have shown through empirical analysis that the approach correlates well with existing measures, while also being different enough that we believe it is measuring different qualitative properties of rankings. The differences can potentially be attributed to various implicit factors that influence user preferences, such as recency, readability,

completeness, and more. Furthermore, the approach works even when the triplet judgments are very far from complete: simply sampling a small set of triplets uniformly at random provided enough information to discriminate between effectiveness of a large set of TREC systems with high correlation to existing measures.

The primary contribution of this work is a new approach to evaluation based on user preferences. Its benefit is that it captures relevance and novelty/diversity, which are qualities that researchers care about, but also allows users to express preferences based on any factors that are important to them—including factors researchers have never thought of or would never think to include in evaluation measures. While it is impossible to prove that any one evaluation framework is better than another, we believe that our experiments, building from small, specific hypotheses about preferences (Section 4.2) to a larger pilot study with users (Section 5.2.2.2) to simulations treating subtopics as the single most important factor in preferences (Section 5.3.1) to a real, large-scale user study (Chapter 6) provide strong evidence that our measures work as described.

7.1 Future Work

In this section, we conclude the dissertation by discussing two avenues of future work. First, we explore the possibility of using our framework to measure the total utility of a system. Second, we focus on learning a novelty function from the user preferences obtained in this work.

7.1.1 Measuring Total Utility of Systems

In 1971, Cooper [64] pointed out the distinction between topical relevance (he referred to this as *logical relevance*) and utility (or usefulness). He argued that while topical relevance is important, the more important question is "how useful is the retrieved information to the user?". For many years IR evaluation persisted with topical relevance in the sense that a document is relevant even if one sentence is on the topic

related to the information need (binary relevance in TREC). Eventually, graded relevance was introduced and shifted the focus of IR evaluation towards utility, although only slightly. In this work, we introduced a triplet framework and validated its potential to estimate the topical relevance and novelty of a document as well as its potential to capture other factors implicitly.

A direction of future work, then, is towards utilizing the triple framework to understand better the factors that influence user preferences. We briefly outline three lines of future work in this direction: (1) eye-tracking study towards deeper investigation factors that influence user preferences; (2) developing algorithms to infer preference judgments thereby reducing human annotation efforts. We briefly discuss them below:

Eye-tracking Experiments

Early use of eye tracing for information retrieval include investigation on how users interact with a ranked list returned by web search engine [76], while user's evaluation styles were investigated by Aula et al. [12]. The field of reading and information processing has made tremendous progress having application in several areas including information retrieval. Three important concepts include *saccades, fixations* and *regression*: *saccades* are rapid eye movements from one point to another; between the saccades, eyes remain relatively still for about 200-300 ms (for silent reading), which is known as *fixations*. Saccades in reading English text is from left to right but about 10-15% of the saccades are right to left (movements back to previously read lines); and they are referred to as *regression*. Early studies discovered several basic facts about eye movements, including: *saccadic suppression* (eye movement during which no information is perceived), *saccade latency* (time taken to initiate eye movements), etc [127, 128, 113, 108].

These studies have been the foundation of several recent works that used eye movements to investigate the cognitive behavior of the users for several information retrieval related tasks. For instance, the average length of forward saccades was used by Buscher et al. [25] as an indicator of relevance. The general theme of these works was

to build a set of features using the eye-tracking data in order to predict the relevance of a document [119, 147, 106, 104, 26, 27, 28].

The above mentioned works considered relevance of a document independently, and the focus was to estimate the relevance of a document rather and not to understand relevance itself. Nevertheless, the techniques developed could be used for investigating the factors that influence user preferences. The saccades and fixations obtained from users annotating the relevance of documents using the triplet framework could provide valuable insights into how inter-document dependencies affect user preferences.

Algorithms to Infer Preference Judgments

Preference judgments are good; their potential to capture various aspects of relevance has been highlighted throughout this work. However, the fact that preference judgments require a large number judgments to be made for even small number of document is a real stumbling block to its widespread use in information retrieval evaluation. We have taken some initial steps in this direction, proposing a graph-based approach to infer pairwise judgments to minimize annotation efforts [45]. There exists a plethora of work in this domain, for example, incomplete pairwise comparison in analytic hierarchy process is a well studied problem in the field of decision sciences [138, 78, 31, 139]. A future direction is to investigate the adaptability of these methods and develop algorithms to reduce the number of pairwise judgments.

7.1.2 Learning to Rank using User Preferences

Another interesting problem is to model user preferences using machine learning approaches to learn a novelty function. Learning to rank approaches are widely used in the research community for several ranking problems. These learning to rank methods can be categorized as *pointwise*, *listwise* or *pairwise*. The pairwise methods [92] in which the learning-to-rank problem is approximated by a classification problem is more suitable to the preference framework explained in this work. The ranker optimizes a

loss function based on preferences to learn a partial ordering of documents with a goal to minimize the average number of inversions in the ranking.

Extensive research efforts have gone into modeling relevance using preference judgments, yet there is little known work on learning to rank models for novelty. Typically, a novelty ranker needs to have information about previously ranked documents to produce a rank list with not only relevant documents but also novel ones. The user preference obtained using AMT accounts for these inter-document effects which provided a more direct measurement of novelty, therefore making it a suitable training data for a novelty ranker. A natural future direction, thus is to model novelty by developing learning to rank methods thereby improving system effectiveness.

BIBLIOGRAPHY

[1] Amazon mechanical turk. http://www.mturk.com.

[2] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*, page 5, New York, New York, USA, 2009. ACM Press.

[3] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. The relationship between IR effectiveness measures and user satisfaction. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, page 773, 2007.

[4] Azzah Al-Maskari, Mark Sanderson, Paul Clough, and Eija Airio. The good and the bad system: does the test collection predict users' effectiveness? In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, page 59, New York, New York, USA, July 2008. ACM Press.

[5] James Allan. *Topic Detection and Tracking: Event-Based Information Organization*. Springer, 2002.

[6] James Allan, Ben Carterette, and J. Lewis. When will information retrieval be "good enough"? In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 433–440. ACM, 2005.

[7] Omar Alonso and Stefano Mizzaro. Using crowdsourcing for TREC relevance assessment. *Information Processing & Management*, February 2012.

[8] Omar Alonso, D.E. Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. In *ACM SIGIR Forum*, volume 42, pages 9–15. ACM, November 2008.

[9] Jaime Arguello, Fernando Diaz, and Jamie Callan. Learning to aggregate vertical results into web search results. In *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, page 201, New York, New York, USA, 2011. ACM Press.

[10] JA Aslam, E Yilmaz, and V Pavlu. A Geometric Interpretation of R-precision and Its Correlation with Average Precision. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 573–574, 2005.

[11] Javed A Aslam, Virgil Pavlu, and Robert Savell. A unified model for metasearch, pooling, and system evaluation. *Proceedings of the twelfth international conference on Information and knowledge management – CIKM'03*, 2003.

[12] Anne Aula, Päivi Majaranta, and Kari-jouko Räihä. Eye-Tracking Reveals the Personal Styles for Search Result Evaluation. In *Proceedings of the 2005 IFIP TC13 international conference on Human-Computer Interaction (INTERACT'05)*, volume 3585, pages 1058–1061, 2005.

[13] David Banks, Paul Over, and Nien-fan Zhang. Blind Men and Elephants : Six Approaches to TREC data. *Information Retrieval*, 1(1-2):7–34, 1999.

[14] Marcia J Bates. Indexing and access for digital libraries and the internet: Human, database, and domain factors. *Journal of the American Society for Information Science*, 49(13):1185–1205, 1998.

[15] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, and Ophir Frieder. Using manually-built web directories for automatic evaluation of known-item retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03*, page 373, New York, New York, USA, July 2003. ACM Press.

[16] Nicholas.J. Belkin, Robert.N. Oddy, and Helen.M. Brooks. ASK for information retrieval: Part I. Background and theory. *Journal of Documentation*, 38(2):61–71, 1982.

[17] Michael Bendersky, Donald Metzler, and W. Bruce Croft. Learning concept importance using a weighted dependence model. *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, page 31, 2010.

[18] Adam Berger and John Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*, pages 222–229, New York, New York, USA, August 1999. ACM Press.

[19] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, May 2003.

[20] David Bodoff and Pu Li. Test theory for assessing IR test collections. In *Proceedings of the 30th annual international ACM SIGIR conference on Research*

and development in information retrieval - SIGIR '07, page 367, New York, New York, USA, July 2007. ACM Press.

[21] Bert Boyce. Beyond Topicality: A Two Stage View of Relevance And The Retrieval Process. *Information Processing & Management*, 18(3):105–109, January 1982.

[22] Bertram C. Brookes. Measurement in information science: Objective and subjective metrical space. *Journal of the American Society for Information Science*, 31(4):248–255, August 1980.

[23] Chris Buckley and Ellen M Voorhees. Evaluating evaluation measure stability. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval – SIGIR'03*, 2000.

[24] Chris Buckley and Ellen M Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, page 25, New York, New York, USA, 2004. ACM Press.

[25] Georg Buscher, Andreas Dengel, Ralf Biedert, and Ludger V. Elst. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM Transactions on Interactive Intelligent Systems*, 1(2):1–30, January 2012.

[26] Georg Buscher, Andreas Dengel, and Ludger van Elst. Eye movements as implicit relevance feedback. In *CHI '08 extended abstracts on Human factors in computing systems*, pages 2991–2996, New York, NY, USA, 2008. ACM Press.

[27] Georg Buscher, Andreas Dengel, and Ludger van Elst. Query expansion using gaze-based feedback on the subdocument level. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, page 387, New York, New York, USA, 2008. ACM Press.

[28] Georg Buscher, Ludger van Elst, and Andreas Dengel. Segment-level display time as implicit feedback. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, page 67, New York, New York, USA, 2009. ACM Press.

[29] Gabriele Capannini, Franco Maria Nardini, Raffaele Perego, and Fabrizio Silvestri. Efficient diversification of web search results. *Proceedings of the VLDB Endowment*, 4(7):451–459, April 2011.

[30] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st*

annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98, pages 335–336, New York, New York, USA, 1998. ACM Press.

[31] Frank J. Carmone, Ali Kara, and Stelios H. Zanakis. A Monte Carlo investigation of incomplete pairwise comparison matrices in AHP. *European Journal of Operational Research*, 102(3):538–553, November 1997.

[32] Claudio Carpineto and Giovanni Romano. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Computing Surveys*, 44(1):1–50, January 2012.

[33] Ben Carterette. An analysis of NP-completeness in novelty and diversity ranking. *Information Retrieval*, 14(1):89–106, December 2010.

[34] Ben Carterette. System effectiveness, user models, and user utility. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, page 903, New York, New York, USA, 2011. ACM Press.

[35] Ben Carterette and James Allan. Incremental Test Collections. In *Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05*, pages 680–687, 2005.

[36] Ben Carterette, James Allan, and Ramesh Sitaraman. Minimal test collections for retrieval evaluation. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, page 268, 2006.

[37] Ben Carterette, Paul N Bennett, David Maxwell Chickering, and T Susan. Here or There Preference Judgments for Relevance. In *Proceedings of the 30th European conference on Advances in information retrieval (ECIR'08)*, pages 16–27, 2008.

[38] Ben Carterette and Praveen Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*, page 1287, New York, New York, USA, 2009. ACM Press.

[39] Ben Carterette, Evgeniy Gabrilovich, Vanja Josifovski, and Donald Metzler. Measuring the reusability of test collections. In *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, page 231, New York, New York, USA, 2010. ACM Press.

[40] Ben Carterette and Rosie Jones. Evaluating Search Engines by Modeling the Relationship Between Relevance and Clicks. In *Proceedings of the 21st Annual*

Conference on Neural Information Processing Systems - NIPS, volume 20, pages 217–224, 2007.

[41] Ben Carterette and Desislava Petkova. Learning a ranking from pairwise preferences. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, page 629, New York, New York, USA, 2006. ACM Press.

[42] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11-16):1623–1640, May 1999.

[43] Praveen Chandar and Ben Carterette. Diversification of search results using webgraphs. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, page 869, New York, New York, USA, 2010. ACM Press.

[44] Praveen Chandar and Ben Carterette. Analysis of Various Evaluation Measures for Diversity. *Proceedings of the 1st International Workshop on Diversity in Document Retrieval at European Conference on Information Retrieval - ECIR'11*, 2011.

[45] Praveen Chandar and Ben Carterette. Using PageRank to infer user preferences. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*, page 1167, New York, New York, USA, 2012. ACM Press.

[46] Praveen Chandar and Ben Carterette. Using preference judgments for novel document retrieval. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*, page 861, 2012.

[47] Praveen Chandar and Ben Carterette. What Qualities Do Users Prefer in Diversity Rankings? In *Proceedings of the 2nd International Workshop on Diversity in Document Retrieval at Web Search and Data Mining Conference - WSDM'12*, 2012.

[48] Praveen Chandar and Ben Carterette. Preference based evaluation measures for novelty and diversity. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13*, page 413, New York, New York, USA, July 2013. ACM Press.

[49] Olivier Chapelle, Shihao Ji, Ciya Liao, Emre Velipasaoglu, Larry Lai, and Su-Lin Wu. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval*, 14(6):572–592, May 2011.

[50] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*, page 621, 2009.

[51] Harr Chen and D.R. Karger. 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*, pages 429–436. ACM, 2006.

[52] Junghoo Cho, Hector Garcia-Molina, and Lawrence Page. Efficient crawling through URL ordering. *Proceedings of the seventh international conference on World Wide Web*, 30(1-7):161–172, April 1998.

[53] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. Overview of the TREC 2009 Web Track. In *Proceedings of the 18th Text REtrieval Conference - TREC'09*, pages 1–9, Gaithersburg, Maryland, November 2009. NIST.

[54] Charles L A Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 75–84. ACM, 2011.

[55] Charles L A Clarke, Nick Craswell, Ian Soboroff, and Ellen M Voorhees. Overview of the TREC 2011 Web Track. In *Proceedings of The Twentieth Text REtrieval Conference - TREC '11*, pages 1–9, 2011.

[56] Charles L. A. Clarke, Nick Craswell, and Ellen M. Voorhees. Overview of the TREC 2012 Web Track. In *Proceedings of 21st Text REtrieval Conference - TREC'12*, 2012.

[57] Charles L A Clarke, Maheedhar Kolla, and Olga Vechtomova. An effectiveness measure for ambiguous and underspecified queries. *Advances in Information Retrieval Theory*, pages 188–199, 2010.

[58] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, page 659, 2008.

[59] CL Clarke and Nick Craswell. Overview of the TREC 2010 Web Track. In *Proceedings of The 19th Text REtrieval Conference - TREC'10*, pages 1–9, 2010.

[60] Cyril W Cleverdon. Report on the Testing and Analysis of an Investigation Into the Comparative Efficiency of Indexing Systems. *ASLIB Cranfield Research Project*, 1962.

[61] Cyril W Cleverdon and J.S. Kidd. Redundancy, Relevance, and Value to the User in The Outputs of Information Retrieval Systems. *Journal of Documentation*, 32(3):159–173, December 1976.

[62] Paul Clough, Mark Sanderson, Murad Abouammoh, Sergio Navarro, and Monica Paramita. Multiple approaches to analysing query diversity. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, page 734, New York, New York, USA, July 2009. ACM Press.

[63] Charles Cole. A theory of information need for information retrieval that connects information to knowledge. *Journal of the American Society for Information Science and Technology*, 62(7):1216–1231, July 2011.

[64] W.S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19–37, June 1971.

[65] GV Cormack and CR Palmer. Efficient construction of large test collections. *Proceedings of the 21st annual*, 1998.

[66] Nick Craswell and David Hawking. Overview of the TREC-2002 Web Track. In *Proceedings of Text REtrieval Conference TREC2002*, pages 1–10, 2003.

[67] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the international conference on Web search and web data mining - WSDM '08*, pages 87–94, New York, New York, USA, February 2008. ACM Press.

[68] Steve Cronen-Townsend and Bruce W. Croft. Quantifying Query Ambiguity. In *Proceedings of the second International Conference on Human Language Technology Research - HLT '02*, pages 104–109. Morgan Kaufmann Publishers Inc., March 2002.

[69] Van Dang and Bruce W. Croft. Term level search result diversification. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13*, page 603, New York, New York, USA, 2013. ACM Press.

[70] Van Dang and W. Bruce Croft. Diversity by Proportionality: An Election-based Approach to Search Result Diversication. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*, page 65, New York, New York, USA, August 2012. ACM Press.

[71] M Eisenberg and C Barry. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of American Society for Information Science*, 39(5):293–300, 1988.

[72] R.A. Fairthorne. Implications of Test Procedures. *Information Retrieval in Action*, pages 109–113, 1963.

[73] William Goffman. On Relevance as a Measure. *Information Storage and Retrieval*, 2(3):201–203, 1964.

[74] Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach for result diversification. In *Proceedings of the 18th international conference on World wide web - WWW '09*, page 381, New York, New York, USA, 2009. ACM Press.

[75] E Graf and L Azzopardi. A methodology for building a patent test collection for prior art search. In *The Second International Workshop on Evaluating Information Access - EVIA '08*, pages 60–71, 2008.

[76] Laura a. Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, page 478, New York, NY, USA, 2004. ACM Press.

[77] Z Gyöngyi, H Garcia-Molina, and Jan Pedersen. Combating web spam with trustrank. *Proceedings of the Thirtieth international conference on Very large databases*, 30:576–587, 2004.

[78] P.T. Harker. Incomplete pairwise comparisons in the analytic hierarchy process. *Mathematical Modelling*, 9(11):837–848, January 1987.

[79] D Harman. Overview of the Fourth Text REtrieval Conference (TREC-4). In *Proceedings of the Fourth TExt Retreival Conference*, pages 1–23, 1996.

[80] Donna K Harman. Overview of the Second Text REtrieval Conference (TREC-2). In *Proceedings of the Second conference on Text retrieval Conference*, pages 1–20, 1993.

[81] Donna K Harman. Overview of the TREC 2002 Novelty Track. In *Proceedings of the Eleventh Text Retrieval Conference - TREC*, volume 2, pages 46–56, 2002.

[82] D Hawking. Challenges in Enterprise Search. In *Proceedings of the Fifteenth Australasian Database Conference - ADC'04*, pages 15–24, 2004.

[83] David Hawking and Stephen Robertson. On Collection Size and Retrieval Effectiveness. *Information Retrieval*, 6(1):99–150, 2003.

[84] Jiyin He, Vera Hollink, and Arjen de Vries. 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 - SIGIR '12*, page 851, New York, New York, USA, August 2012. ACM Press.

[85] Marti A. Hearst and Christian Plaunt. Subtopic structuring for full-length document access. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '93*, pages 59–68, New York, New York, USA, July 1993. ACM Press.

[86] William Hersh, RT Bhuptiraju, and L Ross. TREC 2004 Genomics Track Overview. In *The Fifteenth Text Retrieval Conference*, pages 52–78, 2006.

[87] William Hersh, Chris Buckley, TJ Leone, and D Hickam. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *SIGIR'94*, pages 192 – 201, 1994.

[88] Jian Hu, Gang Wang, Fred Lochovsky, Jian-tao Sun, and Zheng Chen. Understanding user's query intent with wikipedia. In *Proceedings of the 18th international conference on World wide web - WWW '09*, page 471, New York, New York, USA, 2009. ACM Press.

[89] Scott B. Huffman and Michael Hochster. How well does result relevance predict session satisfaction? In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, page 567, 2007.

[90] K Järvelin and J Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '00*, pages 41–48, 2000.

[91] Kalervo Jarvelin and Jaana Kekalainen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[92] T Joachims. Optimizing search engines using clickthrough data. *conference on Knowledge discovery and data mining*, 2002.

[93] Noriko Kando. Evaluation of Information Access Technologies at NTCIR Workshop. In *4th Workshop of the Cross-Language Evaluation Forum, CLEF '03*, pages 29–43, 2003.

[94] PB Kantor and EM Voorhees. The TREC-5 confusion track: Comparing retrieval methods for scanned text. *Information Retrieval*, 2(2-2):165–176, 2000.

[95] R.V. Katter. The influence of scale form on relevance judgments. *Information Storage and Retrieval*, 4(1):1–11, March 1968.

[96] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, September 1999.

[97] M Lalmas and A Tombros. Evaluating XML retrieval effectiveness at INEX. *ACM SIGIR Forum*, 41(1):40 – 57, 2007.

[98] Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01*, pages 120–127, New York, New York, USA, 2001. ACM Press.

[99] Teerapong Leelanupab, Guido Zuccon, and Joemon M Jose. A Query-Basis Approach to Parametrizing Novelty-Biased Cumulative Gain. In *Proceedings of the Third international conference on Advances in information retrieval theory - ICTIR '11*, pages 327–331, 2011.

[100] ME Lesk and Gerard Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4(4):343–359, 1968.

[101] X Li, YY Wang, and A Acero. Learning Query Intent from Regularized Click Graphs. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval -SIGIR '08*, pages 339–346, 2008.

[102] Shangsong Liang, Zhaochun Ren, and Maarten de Rijke. Fusion Helps Diversification. In *Proceedings of the 37th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2014.

[103] Tie-Yan Liu. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, March 2007.

[104] Tomasz D. Loboda, Peter Brusilovsky, and Jöerg Brunstein. Inferring word relevance from eye-movements of readers. In *Proceedings of the 15th international conference on Intelligent user interfaces - IUI '11*, page 175, New York, New York, USA, 2011. ACM Press.

[105] Donald Metzler and W. Bruce Croft. A Markov random field model for term dependencies. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05*, page 472, 2005.

[106] Tristan Miller and Stefan Agne. Attention-based information retrieval using eye tracker data. In *Proceedings of the 3rd international conference on Knowledge capture - K-CAP '05*, page 209, New York, New York, USA, 2005. ACM Press.

[107] Stefano Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, September 1997.

[108] KK Moe and JM Jensen. A qualitative look at eye-tracking for implicit relevance feedback. In *Proceedings of the 2nd International Workshop on Context-Based Information Retrieval*, pages 36–47, 2007.

[109] Alistair Moffat, William Webber, and Justin Zobel. Strategic system comparisons via targeted relevance judgments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, page 375, New York, New York, USA, July 2007. ACM Press.

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

[111] DW Oard, D Soergel, and David Doermann. Building an Information Retrieval Test Collection for Spontaneous Conversational Speech. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48, 2004.

[112] John O'Connor. Retrieval of answer-sentences and answer-figures from papers by text searching. *Information Processing & Management*, 11(5-7):155–164, January 1975.

[113] Takehiko Ohno. EyePrint: Support of Document Browsing with Eye Gaze Trace. In *Proceedings of the 6th international conference on Multimodal interfaces - ICMI '04*, page 16, New York, New York, USA, 2004. ACM Press.

[114] Paul Over. The TREC interactive track: an annotated bibliography. *Information Processing & Management*, 37(3):369–381, May 2001.

[115] Sandeep Pandey and Christopher Olston. User-centric Web crawling. In *Proceedings of the 14th international conference on World Wide Web - WWW '05*, page 401, New York, New York, USA, 2005. ACM Press.

[116] Virgil Pavlu, Peter B Golbus, Javed A Aslam, and Huntington Ave. A Nugget-based Test Collection Construction Paradigm. In *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, pages 1945–1948, 2011.

[117] Virgil Pavlu, Shahzad Rajput, and PB Golbus. IR system evaluation using nugget-based test collections. *Proceedings of the fifth ACM*, page 393, 2012.

[118] Jie Peng, Craig Macdonald, Ben He, Vassilis Plachouras, and Iadh Ounis. Incorporating Term Dependency in the DFR Framework. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, page 843, New York, New York, USA, July 2007. ACM Press.

[119] Michael Pfeiffer. Predicting Text Relevance from Sequential Reading Behavior. In *Proceedings of the NIPS 2005 Workshop on Machine Learning for Implicit Feedback and User Modeling, Whistler, British Columbia, Canada*, pages 25–30, 2005.

[120] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98*, pages 275–281, New York, New York, USA, August 1998. ACM Press.

[121] Filip Radlinski, Paul N. Bennett, Ben Carterette, and Thorsten Joachims. Redundancy, diversity and interdependent document relevance. *ACM SIGIR Forum*, 43(2):46, December 2009.

[122] Filip Radlinski and Susan Dumais. Improving personalized web search using result diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, page 691, New York, New York, USA, August 2006. ACM Press.

[123] Filip Radlinski and Robert Kleinberg. Learning diverse rankings with multi-armed bandits. *on Machine learning*, pages 784–791, 2008.

[124] Filip Radlinski, Martin Szummer, and Nick Craswell. Inferring query intent from reformulations and clicks. In *Proceedings of the 19th international conference on World wide web*, pages 1171–1172, New York, New York, USA, 2010. ACM.

[125] Filip Radlinski, Martin Szummer, and Nick Craswell. Metrics for assessing sets of subtopics. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10*, page 853, New York, New York, USA, July 2010. ACM Press.

[126] Karthik Raman, Pannaga Shivaswamy, and Thorsten Joachims. Online learning to diversify from implicit feedback. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, page 705, New York, New York, USA, August 2012. ACM Press.

[127] K Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372–422, November 1998.

[128] Keith Rayner. *Eye movements and attention in reading, scene perception, and visual search.*, volume 62. August 2009.

[129] Alan M Rees and Douglas G. Schultz. A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching. Final Report to the National Science Foundation. Volume I. September 1967.

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

[131] SE Robertson and S Walker. Okapi at TREC-3. In *Proceedings of 3rd Text REtrieval Conference*, pages 109–126, 1994.

[132] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, January 2004.

[133] Stephen E Robertson. The methodology of information retrieval experiment. In *Information Retrieval Experiment*, pages 9–31. 1981.

[134] Stephen E Robertson and David A Hull. The TREC-9 filtering track final report. In *Proceedings of the Ninth Text REtrieval Conference (TREC-2001)*, pages 25–40, 2001.

[135] Stephen E Robertson and Stephen Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 94*, pages 232–241. Springer-Verlag New York, Inc., August 1994.

[136] Mark E. Rorvig. The simple scalability of documents. *Journal of the American Society for Information Science*, 41(8):590–598, December 1990.

[137] Ronald Rosenfeld. Two Decades Of Statistical Language Modeling: Where Do We Go From Here? *Proceedings of IEEE*, 88(8):1270–1278, 2000.

[138] Thomas L. Saaty. Decision-making with the AHP: Why is the principal eigenvector necessary. *European Journal of Operational Research*, 145(1):85–91, February 2003.

[139] Thomas L Saaty and Luis G Vargas. Inconsistency and rank preservation. *Journal of Mathematical Psychology*, 28(2):205–214, June 1984.

[140] Tetsuya Sakai. Evaluating information retrieval metrics based on bootstrap hypothesis tests. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, pages 525 – 532, 2006.

[141] Tetsuya Sakai. Alternatives to Bpref. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, page 71, New York, New York, USA, July 2007. ACM Press.

[142] Tetsuya Sakai. Evaluating information retrieval metrics based on bootstrap hypothesis tests. *IPSJ Digital Courier*, 3:625–642, 2007.

[143] Tetsuya Sakai, Nick Craswell, Ruihua Song, Stephen Robertson, Z. Dou, and C.Y. Lin. Simple evaluation metrics for diversified search results. In *Proceedings of the 3rd International Workshop on Evaluating Information Access (EVIA)*, 2010.

[144] Tetsuya Sakai, Zhicheng Dou, Ruihua Song, and Noriko Kando. The Reusability of a Diversified Search Test Collection. In *In the Proceedings of the 8th Asia Information Retrieval Societies Conference, AIRS 2012*, pages 26–38, 2012.

[145] Tetsuya Sakai and MP Kato. Click the search button and be happy: evaluating direct and immediate information access. *international conference on Information*, 2011.

[146] Tetsuya Sakai and Ruihua Song. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, pages 1043–1052. ACM, 2011.

[147] Jarkko Saloj, Jaana Simola, Lauri Kovanen, J Salojärvi, and K Puolamäki. Inferring relevance from eye movements: Feature extraction. In *Tech. Rep, Helsinki Univ. of Technology, Publications in Computer and Information Science*, pages 1–23, 2005.

[148] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, November 1975.

[149] Gerard Salton and Chris Buckley. Term Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[150] Gerard Salton and Clement T. Yu. On the construction of effective vocabularies for information retrieval. In *Proceedings of the 1973 meeting on Programming languages and information retrieval - SIGPLAN '73*, pages 48–60, New York, New York, USA, 1973. ACM Press.

[151] Mark Sanderson. Ambiguous queries: test collections need more sense. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, page 499, New York, New York, USA, 2008. ACM Press.

[152] Mark Sanderson. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

[153] Mark Sanderson, M.L. Paramita, Paul Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up? In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 555–562. ACM, 2010.

[154] Mark Sanderson and Justin Zobel. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 162–169, 2005.

[155] Rodrygo L. T. Santos, Jie Peng, Craig Macdonald, and Iadh Ounis. Explicit search result diversification through sub-queries. In *Proceedings of the 32nd European conference on Advances in Information Retrieval - ECIR '10*, volume 5993 of *Lecture Notes in Computer Science*, pages 87–99, March 2010.

[156] Rodrygo L.T. Santos, Craig Macdonald, and Iadh Ounis. Exploiting query reformulations for web search result diversification. *Proceedings of the 19th international conference on World wide web - WWW '10*, page 881, 2010.

[157] Rodrygo L.T. Santos, Craig Macdonald, and Iadh Ounis. Intent-aware search result diversification. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, page 595, 2011.

[158] Iadh Santos, Rodrygo L.T. and Macdonald, Craig and Ounis. Selectively Diversifying Web Search Results. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1179—-1188, 2010.

[159] Tefko Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343, November 1975.

[160] Linda Schamber, Michael B. Eisenberg, and Michael S. Nilan. A re-examination of relevance: toward a dynamic, situational definition. *Information Processing & Management*, 26(6):755–776, January 1990.

[161] Falk Scholer, Diane Kelly, and WC Wu. The Effect of Threshold Priming and Need for Cognition on Relevance Calibration and Assessment. *. . . the 36th international ACM . . .*, 2013.

[162] A Slivkins, F Radlinski, and S Gollapudi. Learning optimally diverse rankings over large document collections. In *Proceedings of the 27th International Conference on Machine Learning- ICML '10*, pages 983–990, 2010.

[163] Catherine L. Smith and Paul B. Kantor. User adaptation: Good Results from Poor Systems. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, page 147, New York, New York, USA, July 2008. ACM Press.

[164] I Soboroff and D Harman. Overview of the TREC 2003 Novelty Track. In *Proceedings of the Twelfth Text Retrieval Conference - TREC*, 2003.

[165] Ian Soboroff. On Evaluating Web Search With Very Few Relevant Documents. In *Proceedings of the 26th Annual International ACM SIGIR Confer- ence on Research and Development in Information Retrieval,*, pages 243–250, 2003.

[166] Ian Soboroff. Overview of the TREC 2004 Novelty Track. In *Proceedings of the Thirteenth Text Retrieval Conference - TREC*, 2004.

[167] Ian Soboroff, Charles Nicholas, and Patrick Cahan. Ranking Retrieval Systems without Relevance Judgments. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 66–73, 2001.

[168] F Song and WB Croft. A General Language Model for Information Retrieval. In *Proceedings of the Eighth International Conference on Information and Knowledge Management - CIKM '99*, pages 316–321, 1999.

[169] Karen Spärck Jones. Report on the need for and provision of an ideal information retrieval test collection. Technical report, British Library Research and Development Report, 1975.

[170] Karen Spärck-Jones, Stephen E Robertson, and Mark Sanderson. Ambiguous Requests: Implications for Retrieval Tests, Systems and Theories. *ACM SIGIR Forum*, 41(2):8–17, December 2007.

[171] Markus Strohmaier, Mark Kröll, and Christian Körner. Intentional query suggestion: making user goals more explicit during search. In *Proceedings of the 2009 workshop on Web Search Click Data - WSCD '09*, pages 68–74, New York, New York, USA, February 2009. ACM Press.

[172] Jean Tague-Sutcliffe. The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, 28(4):467–490, 1992.

[173] Jean Tague-Sutcliffe and J Blustein. A Statistical Analysis of the TREC-3 Data. In *Proceedings of the 3rd Text REtrieval Conference (TREC)*, pages 385–399, 1994.

[174] R Core Team. R: A Language and Environment for Statistical Computing, 2012.

[175] R.G. Throne. The Efficiency of Subject Catalouges and The Cost of Information Search. *Journal of Documentation*, 11(3):130 – 148, 1955.

[176] Ellen M Voorhees. The TREC-8 Question Answering Track Report. In *Proceedings of the Eighth Text Retrieval Conference - TREC*, 1998.

[177] Ellen M Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98*, pages 315–323, New York, New York, USA, August 1998. ACM Press.

[178] Ellen M Voorhees. Overview of TREC-9 Question Answering Track. In *Proceedings of the Ninth Text Retrieval Conference - TREC*, 2000.

[179] Ellen M Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, September 2000.

[180] Ellen M Voorhees. Overview of the TREC 2001 question answering track. In *Proceedings of the Tenth Text Retrieval Conference - TREC*, 2001.

[181] Ellen M Voorhees. Overview of the TREC 2002 Question Answering Track. In *Proceedings of the Eleventh Text Retrieval Conference - TREC*, 2002.

[182] Ellen M Voorhees. The Philosophy of Information Retrieval Evaluation. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer Berlin Heidelberg, 2002.

[183] Ellen M Voorhees. Overview of the TREC 2003 Question Answering Track. In *Proceedings of the Twelfth Text Retrieval Conference - TREC*, pages 1–13, 2003.

[184] Ellen M. Voorhees. Topic set size redux. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, page 806, New York, New York, USA, July 2009. ACM Press.

[185] Ellen M Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 316–323, 2002.

[186] Ellen M Voorhees and Donna Harman. Overview of the Eighth Text REtrieval Conference (TREC-8). In *The Eighth Text Retrieval Conference - (TREC-8)*, pages 1–24. NIST Special Publication, 1999.

[187] Ellen M Voorhees and Donna K Harman. *TREC : Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[188] Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, September 2005.

[189] Jun Wang and Jianhan Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 115–122. ACM, 2009.

[190] Qinglei Wang, Yanan Qian, Ruihua Song, Zhicheng Dou, Fan Zhang, Tetsuya Sakai, and Qinghua Zheng. Mining subtopics from text fragments for a web query. *Information Retrieval*, 16(4):484–503, February 2013.

[191] Xuanhui Wang and ChengXiang Zhai. Learn from web search logs to organize search results. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, page 87, New York, New York, USA, July 2007. ACM Press.

[192] Ryen W White and Dan Morris. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR*, page 255, New York, New York, USA, July 2007. ACM Press.

[193] Emine Yilmaz, Evangelos Kanoulas, and Javed A Aslam. A simple and efficient sampling method for estimating AP and NDCG. *... of the 31st annual international ACM ...*, 2008.

[194] Dawei Yin, Zhenzhen Xue, Xiaoguang Qi, and BD Davison. Diversifying search results with popular subtopics. In *Proceeding of TREC'09*, pages 1–9, 2009.

[195] Yisong Yue and Thorsten Joachims. Predicting diverse subsets using structural SVMs. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 1224–1231, New York, New York, USA, 2008. ACM Press.

[196] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, April 2004.

[197] C.X. Zhai, W.W. Cohen, and John Lafferty. 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*, pages 10–17. ACM, 2003.

[198] Y Zhang, J Callan, and T Minka. Novelty and Redundancy Detection in Adaptive Filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 81–88, 2002.

[199] Wei Zheng, Xuanhui Wang, Hui Fang, and Hong Cheng. An exploration of pattern-based subtopic modeling for search result diversification. In *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries - JCDL '11*, page 387, New York, New York, USA, June 2011. ACM Press.

[200] Wei Zheng, Xuanhui Wang, Hui Fang, and Hong Cheng. Coverage-based search result diversification. *Information Retrieval*, (April), November 2011.

[201] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98*, pages 307–314, 1998.

[202] Guido Zuccon, L. Azzopardi, and K. van Rijsbergen. The quantum probability ranking principle for information retrieval. *Advances in Information Retrieval Theory*, pages 232–240, 2010.

Appendix A

EXPERIMENTAL DATA

A.1 TREC Diversity Collection

TREC Web Track was introduced (*re-introduced*) in 2009 with the goal of studying two tasks: traditional *ad-hoc* task and *diversity task* [53]. For the diversity task, systems were required to return a ranked list of web pages that provide complete coverage for a query, while reducing the amount of redundancy in the ranking. An example provided by the task organizers is given below:

> Given the query "windows" a system might return the Windows update page first, followed by the Microsoft home page, and then a news article discussing the release of Windows 7. Mixed in these results should be pages providing information on doors and windows for homes and businesses.

The TREC-style datasets are available from an annual Text REtrieval Conference (TREC) organized by NIST. A set of retrieval tasks are agreed upon each year, following a two stage process for evaluation. During the first stage, the organizers hire assessor to develop a set of topics for the task and the topics are sent out to research groups participating in the track. A document collection is often agreed upon for the task, research groups are required to search the collection producing a ranked lists of documents for the developed topics and return them to NIST. During the second stage; once the ranked lists from various research (each group can submit multiple ranked lists) are available, the organizers generally pool the top n documents returned by each system for each topic. The pooled set of documents are then judged for relevance by the same assessor that developed the topic and release to the research community. The systems submitted to TREC are also available, these systems represent a variety of

different methods from research groups across the world. They are ideal for studying various research questions with respect to evaluation.

The web track used the ClueWeb09 dataset as the document collection. The collection consists of about 1 billion web pages, comprising approximately 25TB of uncompressed data (5TB compressed) in multiple languages (47% of the documents are in English). The URLs were crawled during January and February of 2009. The dataset is distributed for research purposes only, they can be obtained from Carnegie Mellon University by signing a data license agreement. A smaller subset of the dataset of about 50 million English-language document was created for research groups unable to handle the full collection. All documents are in a highly compressed WARC file format, each WARC file contains several thousand documents. Additionally, a webgraph consisting of about 4,780,950,903 unique URLs and 7,944,351,835 out-links were released as part of the dataset. The webgraph data is a network structure of hyperlinked webpages containing in-links and out-links of each document in the collection.

Topics for the track were created by assessors and the subtopic were obtain from the logs of a commercial search engine with the help of a clustering tool. Topics consisted of a short keyword query and a brief statement describing the user's information need. In order to obtain realistic subtopics, NIST organizers selected candidate subtopics from query logs by identifying queries that frequently co-occurred with the initial target query using random walk algorithm on a bipartite query-document click graph [124]. Then, the candidate subtopics were clustered to reflect multiple aspects of real user needs. Finally, documents were judged with respect to the subtopics on a binary scale and with respect to the topic as whole.

A.2 Newswire Collection

The *Newswire* corpus, was created by Allan et al. [6] to investigate the relationship between system and human performance on a faceted topic retrieval task. The document collection used was created by the Linguistic Data Consortium (LDC), which comprises of news articles from the Agence France Press, Associated Press, Los

Angeles Times, New York Times, and English-edition Xinhua News. Unlike the TREC collection this collection is much smaller in size, there are only about 321,590 articles gathered from October 1, 2003 through March 31, 2004. The documents are in SGML format, containing tags to indicate *title* text and paragraphs in the document.

Topics were created by IR researchers from University of Massachusetts at Amherst by writing a detailed description about their information need. Then, a lead annotator refined these topics to ensure that enough relevant documents existed in the corpus. Queries were manually created for each topic and a total of 130 unique documents were retrieved using a vector space model. Two trained annotators were hired to obtain relevance judgments for all 130 documents for each query (detailed topic descriptions were provided for each query). They were required to identify relevant passage for a given topic and group them by subtopics. The web interface used to obtain judgments aided the grouping of related passage into subtopic by displaying the labels created by the annotators on the side. Therefore, three levels of judgments were obtained: a binary relevance judgment for the document; for each relevant document, a list of subtopics that the document contains; and for each subtopic, a supporting relevant passage in the document.

A.3 Summary of Datasets

Now that we have explained the details of how the datasets were created and the document collections used, we summarize the details of the datasets used in this work. Table A.1 gives an overview of all the datasets used in our experiments. The table includes the data set name, total number of topics, average number of document judged per query, average number of relevance document per query, subtopic range observed. The TREC Diversity task ran from 2009 to 2012, developing relevance judgments for 50 topics each year. Clearly, the TREC dataset is different from the Newswire dataset across a number of characteristics, making it suitable for validation of our experiments. Figure A.1 shows the distribution of subtopic frequency observed in relevant documents for TREC and Newswire datasets. The figure shows that most

Dataset Name	No. of Docs Judged	No. of Relevant Docs	Subtopic Range
TREC09	528.140	98.84	2-8
TREC10	136.52	136.52	2-8
TREC11	1230.40	100.60	2-8
TREC12	1171.70	111.18	2-8
Newswire	127.78	39.43	1-142

Table A.1: An overview of the datasets used in this work. Number of documents judged and number of relevant documents are averaged across queries

relevant documents contain only one or two subtopics, and a decreasing trend can be observed in all of the dataset.

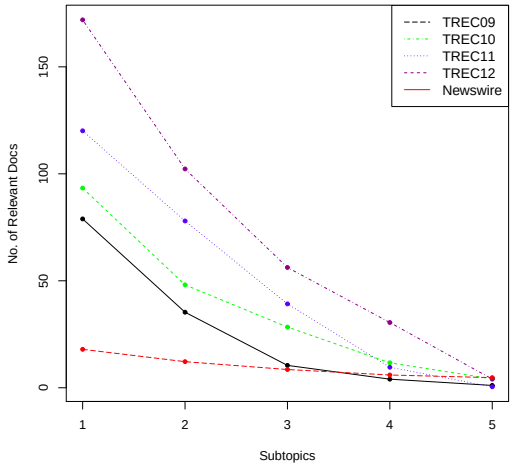

Figure A.1: Number of relevant document for top five popular subtopics averaged across topics.

The newswire corpus was annotated by two different assessors independently, this allows us to compute agreement statistics on the dataset. The agreement about relevance was quite high with 72% of all relevant documents were judged relevant by both assessors, but there was substantial disagreement about the number of subtopics per query, with a difference of 8 subtopics on average. TREC judgments were obtained from a single assessor, thus such agreement statistics could not be obtained.