Foundations and Trends® in Information Retrieval
Vol. 9, No. 1 (2015) 1–90

DOI: 10.1561/1500000040

Search Result Diversification

Rodrygo L. T. Santos
Department of Computer Science
Universidade Federal de Minas Gerais
rodrygo@dcc.ufmg.br

Craig Macdonald
School of Computing Science
University of Glasgow
craig.macdonald@glasgow.ac.uk

Iadh Ounis
School of Computing Science
University of Glasgow
iadh.ounis@glasgow.ac.uk

Contents

Notations

elements

q	A query
a	A relevant query aspect
s	A mined query aspect
d	A document
f	A function (e.g., a ranking function)
r	The rank position of a retrieved document
g_i	The relevance label of the i-th retrieved document

sets

$\mathcal{A}_q$	A set of aspects relevant to a query q
$\mathcal{S}_q$	A set of aspects mined for a query q
$\mathcal{G}_q$	A set of documents relevant for a query q
$\mathcal{R}_q$	A set of documents retrieved for a query q
$\mathcal{D}_q$	A set of documents diversified for a query q

parameters

n	The total number of documents in the corpus
n_q	The number of documents retrieved for the query q
v	The number of unique terms in the corpus
k	The number of aspects underlying a query
κ	An evaluation cutoff
τ	The diversification cutoff
λ	The diversification trade-off

Abstract

Ranking in information retrieval has been traditionally approached as a pursuit of relevant information, under the assumption that the users' information needs are unambiguously conveyed by their submitted queries. Nevertheless, as an inherently limited representation of a more complex information need, every query can arguably be considered ambiguous to some extent. In order to tackle query ambiguity, search result diversification approaches have recently been proposed to produce rankings aimed to satisfy the multiple possible information needs underlying a query. In this survey, we review the published literature on search result diversification. In particular, we discuss the motivations for diversifying the search results for an ambiguous query and provide a formal definition of the search result diversification problem. In addition, we describe the most successful approaches in the literature for producing and evaluating diversity in multiple search domains. Finally, we also discuss recent advances as well as open research directions in the field of search result diversification.

R. L. T. Santos, C. Macdonald and I. Ounis. *Search Result Diversification*.
Foundations and Trends® in Information Retrieval, vol. 9, no. 1, pp. 1–90, 2015.
DOI: 10.1561/1500000040.

1

Introduction

Queries submitted to an information retrieval (IR) system are often ambiguous to some extent. For instance, a user issuing the query "bond" to an IR system could mean the financial instrument for debt security, the classical crossover string quartet "Bond", or Ian Fleming's secret agent character "James Bond". At the same time, the documents retrieved by an IR system for a given query may convey redundant information. Indeed, a user looking for the IMDb page of the James Bond film "Spectre" may be satisfied after observing just one relevant result. Ambiguity and redundancy have been traditionally ruled out by simplifying modelling assumptions underlying most ranking approaches in IR. Nevertheless, in a realistic search scenario, ambiguity and redundancy may render a traditional relevance-oriented ranking approach suboptimal, in terms of subjecting the user to non-relevant results. In this situation, alternative ranking policies should be considered. In this chapter, we provide a historical perspective of relevance-oriented ranking in IR and discuss the challenges posed by ambiguity and redundancy as a motivation for diversifying the search results.

1.1 The Holy Grail of IR

The key challenge faced by an IR system is to determine the *relevance* of a document given a user's query [Goffman, 1964]. The concept of relevance, the holy grail of IR, has been discussed in the fields of information science and retrieval since the 1950s. Despite the rich literature on the subject, relevance per se is still an ill-understood concept [Mizzaro, 1997]. In a practical environment, relevance can span multiple dimensions, related to the topicality and usefulness of the retrieved documents as they are perceived by the target user [Borlund, 2003]. Indeed, relevance is ultimately a prerogative of the user, in which case an IR system can at best estimate it [Baeza-Yates and Ribeiro-Neto, 2011].

Estimating relevance is a challenging task. Indeed, while current search users may have high expectations regarding the quality of the documents returned by a modern web search engine, they often provide the search engine with a rather limited representation of their information need, in the form of a short keyword-based query [Jansen et al., 2000]. Besides understanding the information needs of a mass of users with varying interests and backgrounds, web search engines must also strive to understand the information available on the Web. In particular, the decentralised nature of content publishing on the Web has led to an unprecedentedly large and heterogeneous repository of information, comprising over 30 trillion uniquely addressable documents [Cutts, 2012] in different languages, writing styles, and with varying degrees of authoritativeness and trustworthiness [Arasu et al., 2001].

The enormous size of the Web most often results in an amount of documents matching a user's query that by far exceeds the very few top ranking positions that the user is normally willing to inspect for relevance [Silverstein et al., 1999]. In such a challenging environment, effectively ranking the returned documents, so that the most relevant documents are presented ahead of less relevant ones, becomes of utmost importance for satisfying the information needs of search users [Baeza-Yates and Ribeiro-Neto, 2011]. A standard boolean retrieval is typically insufficient in a web search scenario, in which case more sophisticated approaches can be deployed to produce a ranking of documents likely to be relevant to the user's information need.

1.2 Relevance-oriented Ranking

Probabilistic ranking approaches have been extensively studied in IR as a mechanism to surface relevant information. Although relevance is an unknown variable to an IR system, properties of a query and of a given document may provide evidence to estimate the probability that the document is relevant to the information need expressed by the query. The probability of relevance of a document to a query is central to the well-known probability ranking principle (PRP) in IR [Cooper, 1971, Robertson, 1977, Robertson and Zaragoza, 2009]:

> *"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 practice, as an abstract ranking policy, the PRP does not prescribe how the probability of relevance of a given query-document pair should be estimated. Nonetheless, several probabilistic ranking models have been proposed throughout the years, inspired by the principle. In particular, the literature on probabilistic ranking dates back to 1960, with the seminal work by Maron and Kuhns [1960] on probabilistic indexing and retrieval in a library setting. The field experienced intensive development in the 1970s and 1980s [Cooper, 1971, Harter, 1975a,b, Robertson and Spärck Jones, 1976, Robertson, 1977, Robertson et al., 1981], culminating in some of the most effective ranking functions used by current IR systems [Robertson et al., 1994, 2004, Zaragoza et al., 2004]. Later developments in the field led to effective alternative probabilistic formulations, including statistical language models [Ponte and Croft, 1998, Hiemstra, 1998, Zhai, 2008] and divergence from randomness models [Amati, 2003, 2006].

Despite the relative success attained by the various ranking approaches inspired by the PRP, the development of the principle has been permeated by simplifying modelling assumptions that are often inconsistent with the underlying data [Gordon and Lenk, 1992, Cooper, 1995]. In particular, Gordon and Lenk [1991, 1992] analysed the optimality of the PRP under the light of classical decision and utility theories [von Neumann and Morgenstern, 1944], based upon the costs involved in not retrieving a relevant document as well as in retrieving a non-relevant one. While decision-theoretic costs remain the same for each retrieved document, the utility-theoretic benefit of a relevant document retrieved depends on the previously retrieved relevant documents. In their analysis, Gordon and Lenk [1991] discussed two key modelling assumptions underlying probabilistic ranking approaches:

A1. The probability of relevance is well-calibrated[1] and estimated with *certainty*, with no associated measure of dispersion.

A2. The probability of relevance of a document is estimated *independently* of the other retrieved documents.

According to A1, a document with a higher probability of relevance should always be ranked ahead of a document with a lower probability of relevance, regardless of the confidence of such probability estimates. According to A2, the probability of relevance of a document should be estimated regardless of the probability of relevance of the documents ranked ahead of it. As Gordon and Lenk [1991] demonstrated, the PRP attains the greatest expected utility compared to any other ranking policy under these two assumptions. However, when at least one of these assumptions fails to hold, the principle is suboptimal. In this case, a strict ordering of the retrieved documents by decreasing probability of relevance may not be advisable, and alternative ranking policies should be considered [Gordon and Lenk, 1992]. In general, neither A1 nor A2 are realistic assumptions. In practice, while A1 is challenged by the occurrence of *ambiguity* in the user's query, A2 is challenged by the occurrence of *redundancy* among the retrieved documents.

[1] According to the definition of Gordon and Lenk [1991], a well-calibrated IR system is one that predicts an accurate probability of relevance for each document.

1.3 Ambiguity and Redundancy

Relevance-oriented ranking approaches assume that the users' information needs are unambiguously conveyed by their submitted queries, and that the users' assessment of relevance for a document does not depend on their perceived relevance for the other documents. While such assumptions may have held in the library setting where the early studies of relevance-oriented ranking were conducted [Maron and Kuhns, 1960, Cooper, 1971, Harter, 1975a,b, Robertson, 1977], they do not hold in general [Gordon and Lenk, 1992], and are unlikely to hold in a web search setting, which is permeated with ambiguity and redundancy.

Web search queries are typically short, ranging from two to three terms on average [Jansen et al., 2000]. While short queries are more likely to be ambiguous, every query can be arguably considered ambiguous to some extent [Cronen-Townsend and Croft, 2002]. Nevertheless, in the query understanding literature, query ambiguity is typically classified into three broad classes [Clarke et al., 2008, Song et al., 2009]. At one extreme of the ambiguity spectrum, genuinely *ambiguous queries* can have multiple *interpretations*. For instance, it is generally unclear whether the query *"bond"* refers to a debt security certificate or to Ian Fleming's fictional secret agent character.[2] Next, *underspecified queries* have a clearly defined interpretation, but it may be still unclear which particular *aspect* of this interpretation the user is interested in. For instance, while the query *"james bond"* arguably has a clearly defined interpretation (i.e., the secret agent character), it is unclear whether the user's information need is for books, films, games, etc. Finally, at the other extreme, *clear queries* have a generally well understood interpretation. An example of such queries is *"james bond books"*.

Sanderson [2008] investigated the impact of query ambiguity on web search. In particular, he analysed queries from a 2006 query log of a commercial web search engine that exactly matched a Wikipedia disambiguation page[3] or a WordNet[4] entry. Ambiguous queries from

[2] As a matter of fact, Wikipedia's disambiguation page for *"bond"* lists over 100 possible meanings for this particular entry: `http://en.wikipedia.org/wiki/Bond`.

[3] `http://en.wikipedia.org/wiki/Wikipedia:Disambiguation`

[4] `http://wordnet.princeton.edu`

Wikipedia showed a larger number of senses on average than those from WordNet (7.39 vs. 2.96), with the number of senses per ambiguous query following a power law in both cases. The average length of an ambiguous query was also similar across the two sources, with the predominance of single-word queries. In contrast to previous works, which assumed that multi-word queries were relatively unaffected by ambiguity, he found that ambiguous queries with more than one term were also numerous. Importantly, he observed that ambiguous queries comprised over 16% of all queries sampled from the log. Independent investigations based on click log analyses [Clough et al., 2009] and user studies [Song et al., 2009] also reached the consensual figure that around 16% of all user queries are ambiguous, while many more can be underspecified to some degree. As Sanderson [2008] demonstrated through a simulation, current search systems underperform for such queries.

While ambiguity primarily affects retrieval requests, redundancy is a property of the retrieval results. A document may be considered redundant whenever it conveys information already conveyed by the other documents [Bernstein and Zobel, 2005]. The limitation of assuming that documents are conditionally independent given the query was early recognised. In his note on relevance as a measurable quantity, Goffman [1964] pointed out that *"the relationship between a document and a query is necessary but not sufficient to determine relevance"*. Intuitively, once a document satisfying the user's information need has been observed, it is arguable whether other documents satisfying the same need would be deemed relevant. This intuition has been empirically corroborated in recent years with the analysis of users' browsing behaviour from click logs. Indeed, Craswell et al. [2008] observed that the probability of clicking on a given document diminishes as higher ranked documents are clicked. According to this cascade model, once a user has found the desired information, the need for inspecting further documents is reduced. In practice, the amount of information required to satisfy a user's information need may depend on additional factors. For instance, queries with an informational intent [Welch et al., 2011] as well as those of a controversial nature [Demartini, 2011] may require more than just a single relevant document to satisfy the user.

1.4 Diversity-oriented Ranking

Query ambiguity precludes a clear understanding of the user's actual information need. Wrongly guessing this need may compromise the accuracy of estimating the probability of relevance of any retrieved document. Introducing redundancy may further exacerbate the problem, by promoting more documents related to a potentially wrong information need. Indeed, when the user's actual information need is uncertain, relevance estimations may be misguided, leading to a complete retrieval failure and the abandonment of the query [Chen and Karger, 2006]. In this scenario, a standard relevance-oriented ranking approach is clearly suboptimal, and alternative ranking policies should be considered.

Diversity-oriented ranking has been proposed as a means to overcome ambiguity and redundancy during the search process. Diversifying the search results usually involves a departure from the assumptions that the relevance of a document can be estimated with certainty and independently of the other retrieved documents [Gordon and Lenk, 1991]. Indeed, uncertainty arises naturally from the fact that the probability of relevance is estimated based upon limited representations of both information needs and information items [Turtle and Croft, 1996]. Moreover, it is arguable whether users will still find a given document relevant to their information need once other documents satisfying this need have been observed [Bernstein and Zobel, 2005].

In order to account for both ambiguity and redundancy, a diversity-oriented ranking should not consider the relevance of each document in isolation. Instead, it should consider how relevant the document is in light of the multiple possible information needs underlying the query [Spärck-Jones et al., 2007] and in light of the other retrieved documents [Goffman, 1964]. As a result, the retrieved documents should provide the maximum coverage and minimum redundancy with respect to these multiple information needs [Clarke et al., 2008]. Ideally, the covered information needs should also reflect their relative importance, as perceived by the user population [Agrawal et al., 2009]. In its general form, this is an NP-hard problem [Carterette, 2009], for which an extensive body of research has been devoted in recent years. Discussing such a rich literature is the primary goal of this survey.

1.5 Scope of this Survey

This survey describes several approaches in the literature for the search result diversification problem. In particular, we cover approaches aimed to produce diversity-oriented rankings as well as those aimed at evaluating such rankings. Although our primary focus is on web search, this survey also describes diversification approaches that tackle ambiguity and redundancy in other search scenarios, as well as approaches for related tasks, such as query ambiguity detection and query aspect mining. Outside of the scope of this survey are approaches that seek to promote diversity for purposes other than search, such as text summarisation and event detection and tracking. The notations used uniformly throughout this survey are described in the preface.

The remainder of this survey is organised as follows. In Chapter 2, we provide a comprehensive overview of the search result diversification problem, including a discussion of its NP-hardness. We also describe an approximate polynomial-time solution that underlies most diversification approaches in the literature. These approaches are further organised according to a two-dimensional taxonomy, based upon their adopted aspect representation (implicit or explicit) and their diversification strategy (novelty-based, coverage-based, or hybrid).

In Chapters 3 and 4, we thoroughly describe the most prominent implicit and explicit diversification approaches in the literature, respectively. In both chapters, we focus on the diversification strategy and the ranking objective underlying each approach following the uniform notation introduced in the preface. Throughout these two chapters, we highlight the commonalities and differences among these approaches, and contrast their relative effectiveness as reported in the literature.

In Chapter 5, we describe the evaluation methodology most commonly adopted in the field of search result diversification, which builds upon the availability of benchmark test collections. In particular, we show the overall structure and the core components of a typical test collection for diversity evaluation, and provide a summary of salient statistics of the currently available test collections from TREC and NTCIR. Furthermore, we present multiple alternative evaluation frameworks and detail the evaluation metrics derived from each of them. Finally, we discuss several studies that validate these metrics according to multiple dimensions, including their discriminative power, sensitivity, informativeness, predictive power, optimality, and reusability.

In Chapter 6, we introduce several advanced topics in the field of search result diversification. In particular, we describe approaches proposed for the related tasks of query ambiguity detection and query aspect mining. While the former approaches can be used to selectively adapt the amount of diversification performed for each individual query, the latter can help generate aspect representations that better reflect the possible information needs underlying a user's query. In addition, we describe several diversification approaches introduced for domains other than web search. This includes approaches for diversifying search results in different retrieval domains, such as images, biomedical reports, product reviews and recommendations, as well as for promoting diversity across multiple domains in an aggregated search interface.

Lastly, in Chapter 7, we provide a summary of the materials covered throughout this survey and discuss open research directions in the field of search result diversification. In particular, we highlight open problems related to modelling, estimation, and evaluation of diversification approaches, as a means to foster further research in the field.

2

Search Result Diversification

Ranking in IR has been traditionally approached as a pursuit of relevant information. However, promoting relevance alone may not result in an optimal retrieval effectiveness, particularly in search scenarios that are permeated with ambiguous queries and redundant information items. Search result diversification approaches have been proposed as a means to tackle ambiguity and redundancy. In this chapter, we provide a formal definition of the diversification problem. In addition, we analyse the complexity of the problem and present a classical polynomial-time approximate solution for it. Lastly, we introduce a taxonomy for existing approaches that adhere to this solution.

2.1 The Diversification Problem

Throughout the years, the PRP [Cooper, 1971, Robertson, 1977] has served as a general policy for ranking in IR [Gordon and Lenk, 1991]. However, the development of probabilistic ranking has been permeated by simplifying modelling assumptions that are often inconsistent with the underlying data [Gordon and Lenk, 1992, Cooper, 1995]. In particular, as discussed in §1.2, the PRP assumes that the relevance of a document can be estimated with certainty and independently of the estimated relevance of the other retrieved documents. In practice, neither of these assumptions holds in a realistic scenario. While the first assumption is challenged by ambiguity in the user's query, the second assumption is challenged by redundancy in the search results.

Departing from the aforementioned assumptions requires viewing an ambiguous query as representing not one, but multiple possible information needs [Spärck-Jones et al., 2007]. Under this view, query ambiguity can be tackled by ensuring a high *coverage*[1] of the possible information needs underlying the query among the retrieved documents. In turn, redundancy can be tackled by ensuring that the retrieved documents provide a high *novelty* with respect to their covered needs. Figure 2.1 illustrates rankings with maximum coverage and maximum novelty, derived from an initial relevance-oriented ranking. The figure also illustrates a diversity-oriented ranking, produced by seeking to achieve both coverage and novelty at the same time.

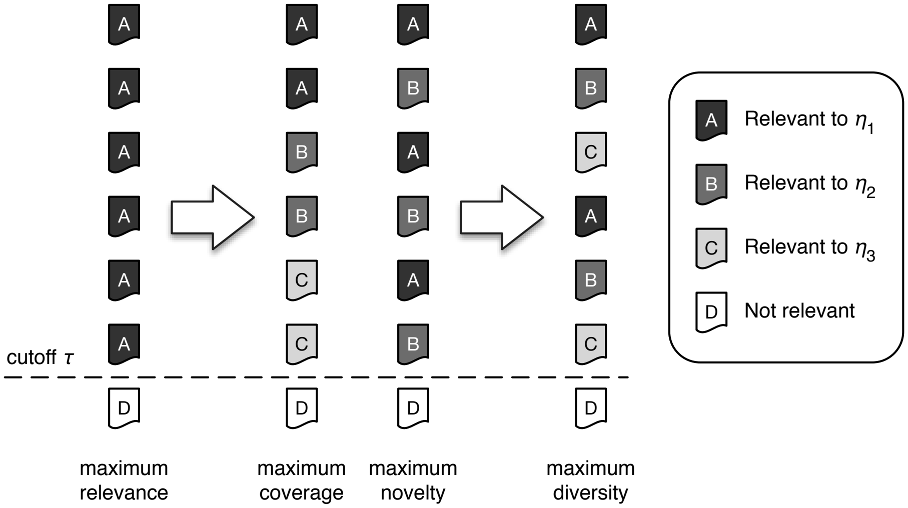

Figure 2.1: Relevance-oriented ranking and the often conflicting goals of diversity-oriented ranking, namely, to attain maximum coverage and maximum novelty.

[1] Clarke et al. [2008] refer to this concept as *"diversity"*. We call it *"coverage"* to emphasise the fact that it is one component of the broader diversification problem.

Coverage and novelty can also be conflicting objectives. Indeed, as illustrated in Figure 2.1, a ranking with maximum coverage may not attain maximum novelty (e.g., although covering all information needs, the ranking may place all documents covering a particular need ahead of documents covering other needs). Conversely, a ranking with maximum novelty may not attain maximum coverage (e.g., although covering each need as early as possible in the ranking, not all possible needs may be covered). As a result, both objectives must be pursued for achieving an optimally diversified ranking. In contrast to a relevance-oriented ranking, which assumes that the user's need is specified with certainty and attempts to retrieve as many relevant documents as possible for this need, a diversity-oriented ranking maximises the chance of retrieving *at least one* relevant document [Chen and Karger, 2006].

The notions of coverage and novelty help provide an informal definition for the search result diversification problem. More formally, let $\mathcal{R}_q$ denote the initial ranking produced for a query q, e.g., by a relevance-oriented ranking approach, such as a best-matching model [Robertson and Zaragoza, 2009], language model [Zhai, 2008], or divergence from randomness model [Amati, 2003].[2] Moreover, following the view of an ambiguous query as representing an ensemble of information needs [Spärck-Jones et al., 2007], let $\mathcal{N}_q$ and $\mathcal{N}_d$ denote the sets of information needs for which the query q and each document $d \in \mathcal{R}_q$ are relevant, respectively. The goal of the search result diversification problem is to find a subset $\mathcal{D}_q \in 2^{\mathcal{R}_q}$, such that:

$$\mathcal{D}_q = \operatorname*{arg\,max}_{\mathcal{D}'_q \in 2^{\mathcal{R}_q}} \left| \cup_{d \in \mathcal{D}'_q} \mathcal{N}_q \cap \mathcal{N}_d \right|, \text{ s.t. } |\mathcal{D}'_q| \leq \tau, \tag{2.1}$$

where $\tau > 0$ is the *diversification cutoff*, denoting the number of top retrieved documents from $\mathcal{R}_q$ to be diversified, and $2^{\mathcal{R}_q}$ is the power set of $\mathcal{R}_q$, comprising all subsets $\mathcal{D}'_q$ of $\mathcal{R}_q$, with $0 < |\mathcal{D}'_q| \leq \tau$, to be considered as candidate permutations of $\mathcal{R}_q$. The permutation with the maximum number of covered information needs up to the cutoff τ is chosen as the optimal diversified ranking $\mathcal{D}_q$.

[2] While diversification could be performed on top of an entire document collection, it is more efficient to start from a smaller set of potentially relevant documents.

2.2 NP-Hardness

As formalised in Equation 2.1, the search result diversification problem is an instance of the maximum coverage problem, a classical NP-hard problem in computational complexity theory [Hochbaum, 1997]. In particular, the maximum coverage problem can be stated as:

> Given a universe of elements $\mathcal{U}$, a collection of potentially overlapping subsets $\mathcal{W} \in 2^{\mathcal{U}}$, and an integer τ, select a set of subsets $\mathcal{M} \subseteq \mathcal{W}$, with $|\mathcal{M}| \leq \tau$, with maximum coverage of the elements from $\mathcal{U}$.

In order to show that the diversification problem is also NP-hard, we can reduce the maximum coverage problem to it [Agrawal et al., 2009]. To this end, we map the universe of elements $\mathcal{U}$ to the possible information needs $\mathcal{N}_q$ underlying the query q. Likewise, we map the collection of candidate subsets $\mathcal{W}$ to the documents in $\mathcal{R}_q$, initially retrieved for q, in which case each document $d \in \mathcal{R}_q$ can be seen as a subset of the information needs $\eta \in \mathcal{N}_q$ for which this document is relevant. As a result, it can be easily verified that a set of subsets $\mathcal{M} \subseteq \mathcal{W}$, with $|\mathcal{M}| \leq \tau$, has maximum coverage of the elements in $\mathcal{U}$ if and only if a permutation $\mathcal{D}_q \subseteq \mathcal{R}_q$, with $|\mathcal{D}_q| \leq \tau$, has maximum diversity with respect to the information needs in $\mathcal{N}_q$.

2.3 Approximate Solution

Since the diversification problem is NP-hard, it has no known efficient exact solution. Instead, we must look for a polynomial-time approximation. An important observation to this end is that the maximisation objective in Equation 2.1 shows a *submodular* structure [Vohra and Hall, 1993]. In particular, given arbitrary sets $\Gamma_1, \Gamma_2 \subseteq \mathcal{U}$, with $\Gamma_1 \subseteq \Gamma_2$, and an element $\gamma \in \mathcal{U} \setminus \Gamma_2$, a set function $f\colon 2^{\mathcal{U}} \to \mathbb{R}$ is called submodular if and only if $f(\Gamma_1 \cup \{\gamma\}) - f(\Gamma_1) \geq f(\Gamma_2 \cup \{\gamma\}) - f(\Gamma_2)$. In other words, adding a new element γ to Γ_1 causes an equal or higher increment in f compared to adding γ to Γ_1's superset Γ_2. Intuitively, a submodular function captures the notion of *decreasing marginal utility*, a fundamental principle in economics [Samuelson and Nordhaus, 2001]. In the context of search result diversification, the marginal utility of selecting a document relevant to an information need diminishes the more this need is satisfied by the documents already selected.

A greedy algorithm can be used to solve the submodular function optimisation in Equation 2.1. As described in Algorithm 2.1, this greedy approach takes as input a query q, the initial ranking $\mathcal{R}_q$ produced for this query, with $|\mathcal{R}_q| = n_q$, and the diversification cutoff $\tau \leq n_q$. As its output, the algorithm produces a permutation $\mathcal{D}_q \subseteq \mathcal{R}_q$, with $|\mathcal{D}_q| = \tau$. Such a permutation is initialised as an empty set in line 1 and iteratively constructed in lines 2–6 of Algorithm 2.1. In line 3, the submodular objective function $f(q, d, \mathcal{D}_q)$ scores each yet unselected document $d \in \mathcal{R}_q \setminus \mathcal{D}_q$ in light of the query q and the documents already in $\mathcal{D}_q$, selected in the previous iterations of the algorithm. The highest scored document, d^*, is then removed from $\mathcal{R}_q$ and added to $\mathcal{D}_q$ in lines 4 and 5, respectively. Finally, in line 7, the produced diverse permutation $\mathcal{D}_q$ of the initial ranking $\mathcal{R}_q$ is returned.

```
**Diversify**$(q, \mathcal{R}_q, \tau)$
1  $\mathcal{D}_q \leftarrow \emptyset$
2  **while** $|\mathcal{D}_q| < \tau$ **do**
3      $d^* \leftarrow \arg\max_{d \in \mathcal{R}_q \setminus \mathcal{D}_q} f(q, d, \mathcal{D}_q)$
4      $\mathcal{R}_q \leftarrow \mathcal{R}_q \setminus \{d^*\}$
5      $\mathcal{D}_q \leftarrow \mathcal{D}_q \cup \{d^*\}$
6  **end while**
7  **return** $\mathcal{D}_q$
```

Algorithm 2.1: Greedy search result diversification.

The asymptotic cost of Algorithm 2.1 is the product of two factors: the cost ϖ_i of evaluating the function f in line 3 at the i-th iteration, and the number Λ_τ of such evaluations required by the algorithm to identify the τ most diverse documents. The unitary cost ϖ_i varies for different approaches, as will be discussed in §2.5. For approaches adhering to the greedy strategy in Algorithm 2.1, the number of evaluations Λ_τ performed up to (and including) the i-th iteration can be modelled as a recurrence relation. In particular, at the first iteration (i.e., $i = 1$), the most diverse document can be trivially selected as the one with the highest estimated relevance to the query, independently of the other documents, since $\mathcal{D}_q = \emptyset$ at this point. At the i-th iteration, with $i > 1$, the function f is evaluated for each document $d \in \mathcal{R}_q \setminus \mathcal{D}_q$, which amounts to a total of $n_q - (i - 1)$ documents. These two observations can be modelled as the base and recursion steps of a first-order linear recurrence, respectively, according to:

$$\Lambda_1 = 0, \tag{2.2}$$

$$\Lambda_i = n_q - i + 1 + \Lambda_{i-1}. \tag{2.3}$$

To obtain the total number of evaluations Λ_τ required to select the τ most diverse documents from $\mathcal{R}_q$, we can iteratively expand the recursion step (Equation 2.3) through telescoping [Cormen et al., 2001], until we arrive at the base step (Equation 2.3), according to:

$$\Lambda_\tau = n_q - \tau + 1 + \Lambda_{\tau-1}, \tag{2.4}$$

$$\Lambda_{\tau-1} = n_q - \tau + 2 + \Lambda_{\tau-2}, \tag{2.5}$$

$$\dots$$

$$\Lambda_2 = n_q - \tau + (\tau - 1) + \Lambda_1. \tag{2.6}$$

Replacing Equation 2.2 into 2.6 and Equations 2.5-2.6 up into Equation 2.4, we can derive a closed form for Λ_τ, as follows:

$$\begin{aligned}\Lambda_\tau &= \sum_{i=2}^{\tau}(n_q - \tau + i) + \Lambda_1 \\ &= \sum_{i=2}^{\tau}(n_q - \tau) + \sum_{i=2}^{\tau} i + 0 \\ &= \frac{1}{2}(2\tau n_q - \tau^2 - 2n_q + \tau). \end{aligned}\tag{2.7}$$

With $\tau \leq n_q$, it follows from Equation 2.7 that $\Lambda_\tau = \mathcal{O}(\tau n_q)$. As $\tau \to n_q$, we have $\Lambda_\tau = \mathcal{O}(n_q^2)$. An important non-approximability result is known for this polynomial-time algorithm, which stems from the submodular structure of the objective function f. In particular, Nemhauser et al. [1978] have shown that such a greedy algorithm achieves an approximation factor of $(1 - 1/e) \approx 0.632$ of the optimal solution to the maximum coverage problem. Feige [1998] has further demonstrated that, for any $\epsilon > 0$, the optimal solution cannot be approximated within a ratio of $(1 - 1/e) + \epsilon$, unless P = NP. This result was independently confirmed under a weaker assumption by Khuller et al. [1999], who proved that no approximation algorithm with ratio better than $(1-1/e)$ exists for the maximum coverage problem, unless NP $\subseteq$ DTIME$(n_q^{\mathcal{O}(\log \log n_q)})$. Alternative formulations of the problem based on local search heuristics have also been shown to perform effectively in practice [Gollapudi and Sharma, 2009, Zuccon et al., 2012]. Nevertheless, given the approximation guarantee offered by Algorithm 2.1, this algorithm underlies most diversification approaches in the literature, as we describe in the next chapters.

2.4 A Taxonomy of Approaches

Several approaches have been proposed in the literature for the search result diversification problem. The vast majority of these approaches differ by how they implement the objective function $f(q, d, \mathcal{D}_q)$ in Algorithm 2.1. Radlinski et al. [2009] broadly classified existing approaches as either *extrinsic* or *intrinsic*. While extrinsic approaches seek to overcome the ambiguity about the user's actual information need, intrinsic approaches aim to avoid redundancy in the search results. Santos et al. [2012b] further refined this classification into a two-dimensional taxonomy, which we adopt in this survey. As shown in Table 2.1, this taxonomy organises existing approaches according to two complementary dimensions: aspect representation and diversification strategy.

An *aspect representation* determines how the information needs underlying a query are represented as multiple *aspects* of this query.[3,4] In particular, an *implicit* aspect representation relies on features belonging to each document in order to model different aspects, such as the terms contained by the documents [Carbonell and Goldstein,

[3] Recall that the information needs underlying an ambiguous query are generally referred to as *query interpretations*, whereas those underlying an underspecified query are referred to as *query aspects* [Clarke et al., 2008]. In this survey, unless otherwise noted, we will refer to both query interpretations and aspects indistinctly.

[4] While both queries and aspects are *representations* of information needs, we find the following distinction helpful: a query is a potentially ambiguous representation of the user's *actual* information need, whereas an aspect is an unambiguous representation of each of the multiple *possible* needs underlying the query.

Table 2.1: Representative diversification approaches in the literature, organised according to two dimensions: diversification strategy and aspect representation.

Diversification strategy	Aspect representation	
	Implicit	Explicit
Novelty	Carbonell and Goldstein [1998] Zhai et al. [2003] Chen and Karger [2006] Zhu et al. [2007] Wang and Zhu [2009] Rafiei et al. [2010] Zuccon and Azzopardi [2010] Gil-Costa et al. [2011, 2013]	Santos et al. [2012b]
Coverage	Radlinski et al. [2008] Carterette and Chandar [2009] He et al. [2011]	Radlinski and Dumais [2006] Capannini et al. [2011]
Hybrid	Yue and Joachims [2008] Slivkins et al. [2010] Santos et al. [2010c] Raman et al. [2012] Zhu et al. [2014] Liang et al. [2014]	Agrawal et al. [2009] Santos et al. [2010b] Dang and Croft [2012]

1998], the clicks they received [Radlinski et al., 2008, Slivkins et al., 2010], or even their different language models [Zhai et al., 2003], topic models [Carterette and Chandar, 2009], or clusters [He et al., 2011] built from the initial document ranking. In turn, an *explicit* aspect representation seeks to directly approximate the possible information needs underlying a query, by relying on features derived from the query itself as candidate aspects, such as different query categories [Agrawal et al., 2009] or query reformulations [Radlinski and Dumais, 2006, Santos et al., 2010b].

Given a particular aspect representation, a *diversification strategy* determines how to achieve the goal of satisfying the multiple aspects underlying a query. In particular, *coverage*-based approaches achieve this goal by directly estimating how well each document covers the multiple aspects of the query, regardless of how well these aspects are covered by the other retrieved documents. Depending on the underlying aspect representation, coverage can be estimated in terms of classification confidence [Agrawal et al., 2009], topicality [Carterette and Chandar, 2009], and relevance [Santos et al., 2010b,c]. In contrast, *novelty*-based approaches directly compare the retrieved documents to one another, regardless of how well they cover different query aspects, in order to promote novel information. For instance, documents can be compared in terms of content dissimilarity [Carbonell and Goldstein, 1998], divergence of language models [Zhai et al., 2003], or relevance score correlation [Rafiei et al., 2010, Wang and Zhu, 2009, Kharazmi et al., 2014]. Coverage and novelty are related to the notions of extrinsic and intrinsic diversity, respectively, as discussed by Radlinski et al. [2009]. Indeed, while a coverage-based strategy primarily focuses on resolving query ambiguity, a novelty-based strategy focuses on avoiding redundancy in the search results. In the remainder of this chapter, we contrast novelty-based, coverage-based, and hybrid diversification strategies in terms of their computational complexity.

2.5 Complexity Analysis

As discussed in §2.2, the asymptotic cost of Algorithm 2.1 is the product of the cost ϖ_i of evaluating the function $f(q, d, \mathcal{D}_q)$ at the i-th iteration, and the number Λ_τ of such evaluations required to identify the τ most diverse documents. While the cost of each evaluation depends on the adopted aspect representation, the number Λ_τ of required evaluations depends on the adopted diversification strategy.

For diversification approaches that attempt to tackle redundancy, such as novelty-based and hybrid approaches, a total of $\Lambda_\tau = \mathcal{O}(\tau n_q)$ evaluations must be performed, as discussed in §2.2. As highlighted in Table 2.1, the vast majority of novelty-based approaches as well as some hybrid approaches adopt an implicit aspect representation, typically comprising the space of unique terms in a document corpus. For such approaches, at the i-th iteration, an evaluation of the objective function $f(q, d, \mathcal{D}_q)$ would have a cost $\varpi_i \propto v(i-1)$, where v is the number of unique terms in the lexicon. Nonetheless, in reality, the function f must only be evaluated with respect to the last document added to $\mathcal{D}_q$ (as opposed to the entire set $\mathcal{D}_q$), since the yet unselected documents in $\mathcal{R}_q \setminus \mathcal{D}_q$ would have already been compared to the documents added to $\mathcal{D}_q$ in the previous iterations. As a result, the total cost incurred by a novelty-based diversification approach can be expressed as $\sum_{i=1}^{\Lambda_\tau} \varpi_i = \sum_{i=1}^{\mathcal{O}(\tau n_q)} v = \mathcal{O}(v \tau n_q)$. For approaches that adopt an explicit aspect representation, the asymptotic cost can be expressed as $\mathcal{O}(k \tau n_q)$, where k is the total number of represented aspects. Compared to the cost $\mathcal{O}(v \tau n_q)$ incurred by implicit approaches, explicit approaches are generally more efficient, since typically $k \ll v$. On the other hand, regarding the aspect representation itself, the identification of implicit, document-driven aspects can be performed offline at indexing time. In contrast, explicit, query-driven aspects generally require online processing at querying time.

For coverage-based approaches, which do not account for dependences between the retrieved documents, the function f is evaluated only once for each of the n_q documents to be diversified. As a result, these approaches perform a total of $\Lambda_\tau = \mathcal{O}(n_q)$ evaluations. Although such an independence assumption breaks the effectiveness guarantees offered by the greedy approximation, it improves the efficiency of the resulting diversification. In particular, while novelty-based approaches evaluate the objective function $f(q, d, \mathcal{D}_q)$ a total of $\mathcal{O}(\tau n_q)$ times, only $\mathcal{O}(n_q)$ evaluations are required by coverage-based approaches. The cost of a single evaluation, in turn, depends on the total number of represented aspects k, i.e., $\varpi_i = \mathcal{O}(k)$. Similarly to the analysis conducted for novelty-based and hybrid approaches, we can express the total cost incurred by coverage-based approaches as $\sum_{i=1}^{\Lambda_\tau} \varpi_i = \sum_{i=1}^{\mathcal{O}(n_q)} k = \mathcal{O}(k n_q)$. In Chapters 3 and 4, we will describe the most representative approaches in each of these families.

3

Implicit Diversification

Search result diversification was originally tackled from a document-oriented perspective. In particular, the first diversification approaches proposed in the literature aimed at identifying documents that carried different information from those documents already seen by the user. As illustrated in Figure 3.1, these approaches implicitly assumed that different documents would more likely satisfy different information needs. The key challenge then became to appropriately model document differences. In this chapter, we describe the most representative diversification approaches in the literature that adopt an implicit representation of query aspects. As discussed in §2.4, depending on their adopted diversification strategy, we further organise these approaches as novelty-based, coverage-based, or hybrid.

3.1 Novelty-based Approaches

Most implicit diversification approaches in the literature seek to promote novelty as their sole diversification strategy. By doing so, their aim is to infer differences between the retrieved documents in order to demote those with redundant contents. The first novelty-

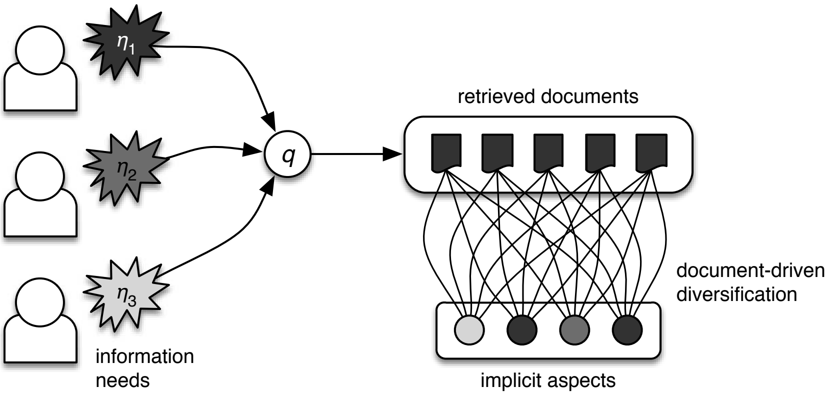

Figure 3.1: Schematic view of implicit diversification approaches.

based diversification approach in the literature was introduced by Carbonell and Goldstein [1998], with applications to text retrieval and summarisation. In particular, their maximal marginal relevance (MMR) method scored a candidate document $d \in \mathcal{R}_q \setminus \mathcal{D}_q$ as the document's estimated relevance with respect to the query q, discounted by the document's maximum similarity with respect to the already selected documents in $\mathcal{D}_q$, according to:

$$f_{\mathrm{MMR}}(q, d, \mathcal{D}_q) = \lambda f_1(q, d) - (1 - \lambda) \max_{d_j \in \mathcal{D}_q} f_2(d, d_j), \tag{3.1}$$

where $f_1(q, d)$ and $f_2(d, d_j)$ estimate the relevance of the document d with respect to the query q and its similarity to the documents already selected in $\mathcal{D}_q$, respectively. A balance between relevance (i.e., f_1) and redundancy (i.e., $\max f_2$, the opposite of novelty) is achieved through an appropriate setting of the linear combination parameter λ.

Inspired by the formulation of MMR, Zhai et al. [2003] proposed a novelty-based diversification approach within a risk minimisation (RM) framework for language modelling [Zhai and Lafferty, 2006]. In particular, given a query q and a candidate document d, their proposed approach estimated the score of the document model θ_d with respect to the query model θ_q, as well as with respect to a reference model $\theta_{\mathcal{D}_q}$, comprising the documents already selected, according to:

$$f_{\mathrm{RM}}(q, d, \mathcal{D}_q) = f_1(\theta_q, \theta_d)(1 - \lambda - f_2(\theta_d, \theta_{\mathcal{D}_q})), \tag{3.2}$$

where $f_1(\theta_q, \theta_d)$ quantifies the risk of returning documents with a language that does not fit the query language, as quantified by the Kullback-Leibler divergence between the θ_q and θ_d language models. Six methods were proposed in order to estimate $f_2(\theta_d, \theta_{\mathcal{D}_q})$, based on either the divergence between θ_d and $\theta_{\mathcal{D}_q}$ or a mixture of the reference model $\theta_{\mathcal{D}_q}$ and an English background model. Similarly to Equation 3.1, the parameter λ controls the penalisation of redundancy.

A related risk-aware approach was introduced by Chen and Karger [2006]. In particular, they argued that maximising the probability of relevance could lead to a complete retrieval failure when ranking under uncertainty. Instead, they proposed to maximise the chance of retrieving at least one relevant document. To this end, they instantiated the objective function in Algorithm 2.1 to estimate the conditional relevance (CR) of a document d, under the assumption that none of the already selected documents $\mathcal{D}_q$ were relevant, such that:

$$f_{\text{CR}}(q, d, \mathcal{D}_q) = p(g_{r(d)} \mid \bar{g}_1, \cdots, \bar{g}_{|\mathcal{D}_q|}, d_1, \cdots, d_{|\mathcal{D}_q|}, d), \tag{3.3}$$

where $r(d)$ denotes the ranking position of document d, and g_i and $\bar{g}_i$ denote the events in which the document at the i-th position is relevant and non-relevant, respectively. Intuitively, this formulation promotes novel documents by considering the already selected documents as a form of negative relevance feedback [Wang et al., 2008].

Wang and Zhu [2009] introduced a diversification approach[1] inspired by the portfolio theory in finance [Markowitz, 1952]. In particular, the selection of documents for a ranking involves a fundamental risk, namely, that of overestimating the relevance of individual documents, analogously to the risk involved in selecting financial assets (e.g., stocks) for an investment portfolio. In both the finance and the retrieval scenarios, diversifying the selected items can maximise the expected return (mean) while minimising the involved risk (variance) of a selection. Accordingly, Wang and Zhu [2009] proposed a mean-variance analysis (MVA) diversification objective, according to:

$$f_{\text{MVA}}(q, d, \mathcal{D}_q) = \mu_d - b\, w_i\, \sigma_d^2 - 2\, b\, \sigma_d \sum_{d_j \in \mathcal{D}_q} w_j\, \sigma_{d_j}\, \rho_{d,d_j}, \tag{3.4}$$

[1] A very similar approach was proposed independently by Rafiei et al. [2010].

where μ_d and σ_d^2 are the mean and variance of the relevance estimates associated with document d, respectively, with the summation component estimating the redundancy of d in light of the documents in $\mathcal{D}_q$. Documents are compared in terms of the Pearson's correlation ρ_{d,d_j} of their term vectors. The weight w_i assigns a discount to the document at the i-th ranking position. A balance between relevance, variance, and redundancy is achieved with the parameter b.

Building upon the theory of quantum mechanics [van Rijsbergen, 2004], Zuccon and Azzopardi [2010] proposed the quantum probability ranking principle (QPRP). In contrast to the classic PRP [Cooper, 1971, Robertson, 1977], the QPRP prescribes that not only the estimated relevance of each document should be considered as a ranking criterion, but also how it interferes with the estimated relevance of the other documents. In particular, in the quantum formalism, interference refers to the effect of an observation on subsequent observations. This notion was quantified into the following objective:

$$f_{\text{QPRP}}(q, d, \mathcal{D}_q) = p(\mathcal{G}_q|q, d) + \sum_{d_j \in \mathcal{D}_q} \varrho_{d,d_j}, \tag{3.5}$$

where $p(\mathcal{G}_q|q, d)$ denotes the probability of observing the relevant set $\mathcal{G}_q$, given the query q and the document d, which corresponds to the classic formulation of the PRP. The estimation of the interference ϱ_{d,d_j} between d and each document $d_j \in \mathcal{D}_q$ involves operations with complex numbers. In practice, it can be approximated as $\varrho_{d,d_j} \approx -2\sqrt{p(\mathcal{G}_q|q, d)}\sqrt{p(\mathcal{G}_q|q, d_j)}\, f(d, d_j)$, where $f(d, d_j)$ can be any function measuring the similarity between the two documents.

Zhu et al. [2007] approached the diversification problem as an absorbing random walk (ARW) with transition probabilities $p_{ij} = (1 - \lambda)\, p(d_j|q) + \lambda\, p(d_j|d_i)$, where $p(d_j|q)$ and $p(d_j|d_i)$ denoted the estimated relevance of d_j and its similarity to d_i, respectively, with the parameter λ balancing between the two. An absorbing random walk is a Markov chain with reachable absorbing states i, such that $p_{ij} = 1$ if $i = j$, and 0 otherwise [Kemeny and Snell, 1960]. In their formulation, each selected document $d_j \in \mathcal{D}_q$ was represented as an absorbing state, in which case candidate documents were scored according to:

$$f_{\mathrm{ARW}}(q, d, \mathcal{D}_q) = \vartheta(d, \mathcal{D}_q), \tag{3.6}$$

where $\vartheta(d, \mathcal{D}_q)$ denotes the expected number of visits to document d before absorption by the states in $\mathcal{D}_q$. While this computation would incur an expensive inversion of the underlying transition matrix at every iteration, in practice, such an inversion can be computed only once and subsequently reused in order to update the portion of the matrix corresponding to the states in $\mathcal{R}_q \setminus \mathcal{D}_q$ [Woodbury, 1950].

Gil-Costa et al. [2011, 2013] explored the properties of the metric space induced from the ranking produced for a query in order to identify novel documents. To this end, they deployed different techniques to partition the initial ranking $\mathcal{R}_q$ into zones $\mathcal{Z}_q$, with each zone comprising documents similar to each other and dissimilar from documents in the other zones. Since $|\mathcal{Z}_q| \ll |\mathcal{D}_q|$, they were able to drastically reduce the number of document comparisons required to promote novelty, by comparing each candidate document $d \in \mathcal{R}_q \setminus \mathcal{D}_q$ to each identified zone centre $z \in \mathcal{Z}_q$, instead of all previously selected documents $d_j \in \mathcal{D}_q$. While such centres could be directly returned as a diverse selection of documents, they introduced a scoring function to perform what they called a sparse spatial selection diversification (SSSD):

$$f_{\mathrm{SSSD}}(q, d, \mathcal{D}_q) = (1 - \lambda)\, f_1(q, d) + \lambda\Big(1 - \max_{z \in \mathcal{Z}_q} f_2(d, z)\Big), \tag{3.7}$$

where $f_1(q, d)$ and $f_2(d, z)$ estimate the relevance of d to the query q and its similarity—as given by a metric distance—to each zone centre z, with the parameter λ controlling the trade-off between the two scores.

3.2 Coverage-based Approaches

A few implicit diversification approaches that adopt a coverage-based strategy have recently been proposed. Rather than directly comparing the retrieved documents to one another, as novelty-based approaches would do, these coverage-based approaches seek to identify a reasonable representation of the aspects underlying a query from the top retrieved documents themselves. For instance, Radlinski et al. [2008] proposed an online learning approach to maximise the coverage of clicked documents. Their intuition was that users with different information needs would click on different documents for the same query. In their formulation, the choice of the next document to be selected for a query was seen as a multi-armed bandit (MAB) problem [Berry and Fristedt, 1985]. A MAB models the process of selecting one of many possible strategies or “arms”, trading off the exploitation of existing knowledge and the acquisition (or exploration) of new knowledge. In the context of diversification, each candidate document $d \in \mathcal{R}_q \setminus \mathcal{D}_q$ for the next ranking position of a query q was considered as an “arm”, with existing knowledge $\mu_d^{(i)}$ at time i denoting the likelihood of the document being clicked when ranked at that position for the query. Precisely, their ranked-armed bandits (RAB) objective can be described as:

$$f_{\mathrm{RAB}}(q, d, \mathcal{D}_q) = \begin{cases} 1 & \text{if } d = \mathrm{MAB}_j(\mathcal{R}_q, \mu_\bullet^{(i)}), \\ 0 & \text{otherwise}, \end{cases} \tag{3.8}$$

where $\mathrm{MAB}_j(\mathcal{R}_q, \mu_\bullet^{(i)})$ is a MAB instance trained to select a document $d^* \in \mathcal{R}_q$ for the j-th ranking position, with $j = |\mathcal{D}_q| + 1$, balancing exploration and the exploitation of the expected reward $\mu_{d^*}^{(i)}$ at time i.

Carterette and Chandar [2009] proposed a probabilistic approach for maximising the coverage of multiple “facets”, denoting different aspects of a query. Such facets were generated by constructing either relevance models [Lavrenko and Croft, 2001] or topic models [Blei et al., 2003] from the top retrieved documents for the query. Three strategies were proposed to re-rank the initially retrieved documents $\mathcal{R}_q$ with respect to their coverage of the identified facets. In particular, the best performing of these strategies selected the highest scored document d for each facet $z \in \mathcal{Z}_q$ in a round-robin fashion. Such a facet modelling (FM) approach can be formalised into the following objective:

$$f_{\mathrm{FM}}(q, d, \mathcal{D}_q) = \begin{cases} p(d|q) & \text{if } \exists z_i \in \mathcal{Z}_q \mid p(d|z_i) > 0 \wedge i = |\mathcal{D}_q| \bmod |\mathcal{Z}_q|, \\ 0 & \text{otherwise}, \end{cases} \tag{3.9}$$

where $\mathcal{Z}_q$ is the set of facets identified for the query q and $p(d|z_i)$ denotes the likelihood of observing each document d given the facet

$z_i \in \mathcal{Z}_q$. The modulus operation ensures a round-robin selection from a total of $|\mathcal{Z}_q|$ facets. Since the probabilities $p(d|z_i)$ are not comparable across the various facets, the documents selected in the round-robin process are ultimately ordered by their likelihood given the query q.

A similar coverage-based approach was investigated by He et al. [2011]. In particular, they proposed to partition the documents initially retrieved for a query into non-overlapping clusters using topic modelling [Blei et al., 2003]. In their approach, each cluster $c \in \mathcal{Z}_q$ received a score $p(c|q)$, given by the cluster's likelihood of generating the query q. As a result, the diversification problem was reduced to the task of selecting documents with a high coverage of highly scored clusters. Of the selection strategies investigated, a weighted round-robin (WRR) selection performed the best. This selection strategy can be formalised according to the following objective:

$$f_{\text{WRR}}(q, d, \mathcal{D}_q) = \begin{cases} p(d|q) & \text{if } \exists c_i \in \mathcal{Z}_q \mid d \in c_i \wedge i = |\mathcal{D}_q| \bmod |\mathcal{Z}_q| \\ & \text{s.t. } p(c_1|q) \geq p(c_2|q) \geq \cdots \geq p(c_{|\mathcal{Z}_q|}|q), \\ 0 & \text{otherwise,} \end{cases} \tag{3.10}$$

where the probability $p(c_i|q)$ imposes a total ordering over the clusters $c_i \in \mathcal{Z}_q$, essentially biasing the selection towards highly scored clusters.

An alternative approach to identify related documents was introduced by Kharazmi et al. [2014]. Instead of exploiting textual features of the retrieved documents, they proposed to leverage the relevance scores initially assigned to these documents to induce "relevance clusters". In particular, they hypothesised that documents with similar relevance scores to already selected documents should be demoted in favour of documents with more dissimilar scores. Accordingly, they proposed to score each retrieved document d by accounting for both the document's original ranking position as well as its score difference with respect to the document ranked immediately ahead of it. Their proposed score difference (SD) objective can be expressed as:

$$f_{\text{SD}}(q, d, \mathcal{D}_q) = \frac{1}{r(d, \mathcal{R}_q)} + \frac{1}{r(d, \mathcal{R}_d)}, \tag{3.11}$$

where $r(d, \mathcal{R}_q)$ and $r(d, \mathcal{R}_d)$ are the positions of d in the original ranking $\mathcal{R}_q$ and in the ranking $\mathcal{R}_d$ induced by score differences, respectively.

3.3 Hybrid Approaches

Hybrid diversification approaches leveraging an implicit aspect representation have also been proposed in recent years. For instance, Yue and Joachims [2008] proposed a hybrid approach within the framework of supervised machine learning. As training data, they considered a pair $(\mathcal{R}_{q_i}, \mathcal{N}_{q_i})$ for each query q_i, where $\mathcal{R}_{q_i}$ and $\mathcal{N}_{q_i}$ denote the initial ranking and the manually labelled information needs possibly underlying q_i, respectively. Since the actual needs $\mathcal{N}_{q_i}$ are unknown in a real scenario, these were implicitly represented using the words covered by each retrieved document. For learning a function f to identify a set $\mathcal{D}_{q_i} \subseteq \mathcal{R}_{q_i}$ with maximum coverage of $\mathcal{N}_{q_i}$, they employed structural support vector machines [Tsochantaridis et al., 2005]. In particular, their weighted word coverage (WWC) approach considered linear functions f, parametrised by a weight vector $\mathbf{w}$, such that:

$$f_{\text{WWC}}(q, d, \mathcal{D}_q) = \mathbf{w}^T \mathbf{\Phi}(\mathcal{R}_q, \mathcal{D}_q \cup \{d\}), \tag{3.12}$$

where the feature extractor $\mathbf{\Phi}(\mathcal{R}_q, \mathcal{D}_q \cup \{d\})$ measures the extent to which the words in $\mathcal{R}_q$ are covered by each candidate selection $\mathcal{D}_q \cup \{d\}$.

Another hybrid diversification approach that leveraged an implicit aspect representation was proposed by Slivkins et al. [2010] within the MAB framework. In particular, they extended the click coverage maximisation approach introduced by Radlinski et al. [2008] and formalised in Equation 3.8 in order to account for the context in which clicks are observed. To this end, they proposed to condition the expected reward $\mu^{(i)}_{d|\mathcal{D}_q}$ of each document d at time i on the documents $\mathcal{D}_q$ selected ahead of d in previous iterations of Algorithm 2.1. This notion can be formalised into the following diversification objective, denoted ranked context bandits (RCB):

$$f_{\text{RCB}}(q, d, \mathcal{D}_q) = \begin{cases} 1 & \text{if } d = \text{MAB}_j(\mathcal{R}_q, \mu^{(i)}_{\bullet|\mathcal{D}_q}), \\ 0 & \text{otherwise,} \end{cases} \tag{3.13}$$

where the instance $\text{MAB}_j(\mathcal{R}_q, \mu^{(i)}_{\bullet|\mathcal{D}_q})$ selects a document $d^* \in \mathcal{R}_q$ for the j-th ranking position, with $j = |\mathcal{D}_q| + 1$, in a similar fashion to Equation 3.8. However, in contrast to Equation 3.8, Slivkins et al. [2010] used the conditional reward $\mu^{(i)}_{d^*|\mathcal{D}_q}$ at time i, by correlating the clicks on d^* to those observed for the documents $d_j \in \mathcal{D}_q$. To reduce the number of required correlation computations, they modelled the reward function $\mu_\bullet$ as a Lipschitz-continuous function in the metric space induced by the documents in $\mathcal{R}_q$ [Searcóid, 2006], which dramatically improved the efficiency of the proposed approach.

A related supervised machine learning approach was introduced by Raman et al. [2012], also in an online learning setting. In particular, at a given time i, their approach presented the user with a diverse ranking $\mathcal{D}_q$, produced by the following objective function:

$$f_{\text{DP}}(q, d, \mathcal{D}_q) = \mathbf{w}_i^T \boldsymbol{\Phi}(\mathcal{R}_q, \mathcal{D}_q \cup \{d\}), \tag{3.14}$$

where $\mathbf{w}_i$ denotes the weight vector learned by a diversification perceptron (DP), based upon the evidence accumulated up to time i, and $\boldsymbol{\Phi}(\mathcal{R}_q, \mathcal{D}_q \cup \{d\})$ is defined in terms of word coverage, similarly to Equation 3.12. In order to update the weight vector $\mathbf{w}_i$, the feedback received from the user in the form of pairwise preferences is used to produce an improved (in expectation) ranking $\hat{\mathcal{D}}_q$. In particular, the updated vector is defined as $\mathbf{w}_{i+1} = \mathbf{w}_i + \boldsymbol{\Phi}(\mathcal{R}_q, \hat{\mathcal{D}}_q) - \boldsymbol{\Phi}(\mathcal{R}_q, \mathcal{D}_q)$.

Zhu et al. [2014] introduced an alternative learning approach, which modelled both relevance-oriented features of each individual document as well as diversity-oriented features conveying each document's pairwise distances to higher ranked documents. Given training instances of the form $(\mathcal{R}_{q_i}, \mathcal{N}_{q_i}, \mathcal{M}_{q_i})$ for each training query q_i, with $\mathcal{R}_{q_i}$ and $\mathcal{N}_{q_i}$ defined as above and the tensor $\mathcal{M}_{q_i}$ representing various distances between every pair of documents in $\mathcal{R}_{q_i}$, they performed a gradient descent optimisation [Friedman, 2001] to minimise a loss function based on the Plackett-Luce model [Marden, 1996]. In particular, their relational learning to rank (RLTR) objective was defined as:

$$f_{\text{RLTR}}(q, d, \mathcal{D}_q) = \mathbf{w}_R^T \boldsymbol{\Phi}(d) + \mathbf{w}_D^T \boldsymbol{\Phi}(\mathcal{M}_q^{d,\mathcal{D}_q}), \tag{3.15}$$

where the vectors $\mathbf{w}_R$ and $\mathbf{w}_D$ are learned to weigh relevance- and diversity-oriented features, respectively. As to the latter, they considered the minimum, average, or maximum distance between d and the documents already selected in $\mathcal{D}_q$, according to dissimilarity functions based on the body, title, anchor text, URLs, hyperlinks, topics, and categories of each pair of documents, as encoded in the tensor $\mathcal{M}_q^{d,\mathcal{D}_q}$.

Liang et al. [2014] observed that fusing multiple rankings produced by independent approaches helps improve diversification. Accordingly, they proposed a diversified data fusion (DDF) model, which aims to promote documents from a fused ranking with a proportional coverage of multiple topics, derived from the document contents as well as their fusion scores. Inspired by the PM-2 model, described in Equation 4.7, Liang et al. [2014] formalised the DDF model according to:

$$f_{\text{DDF}}(q, d, \mathcal{D}_q) = \sum_{z \in \mathcal{Z}_q} b_z \, \zeta(z|q) \, p(d|q, z), \tag{3.16}$$

where $\zeta(z|q)$ is a quotient that determines how much the diversified ranking $\mathcal{D}_q$ already covers the topic z, whereas $p(d|q, z)$ determines how much the document d covers this topic. The latter quantity is further computed as $p(d|q, z) \approx \frac{p(z|q,d) \, f_{\text{CS}}(q,d)}{p(z|q)}$, with $p(z|q, d)$ and $p(z|q)$ estimated via topic modelling [Blei et al., 2003] and $f_{\text{CS}}(q, d)$ estimated by the CombSUM data fusion method [Fox and Shaw, 1993]. Finally, for a given parameter λ, set through training, b_z is set as $b_z = \lambda$ if z is the topic with highest quotient or $b_z = 1 - \lambda$ otherwise.

3.4 Summary

In this chapter, we have surveyed implicit approaches for search result diversification. In particular, these approaches have the longest history in the literature, stemming from research on identifying novel sentences for text summarisation [Carbonell and Goldstein, 1998]. In common, implicit approaches rely on document properties as proxies for representing the information needs covered by each document. Such document properties include terms, clicks, relevance scores, topic models, and clusters. Despite their intuitiveness, these approaches have been shown to underperform in a standard web search result diversification scenario, regardless of their choice of aspect representation [Santos et al., 2012b]. Indeed, seeking novelty—the most popular strategy among implicit approaches—has been shown to yield highly inconsistent improvements across different queries. In contrast, seeking coverage—a strategy popularised by explicit diversification approaches, as introduced in the next chapter—has been shown to be less risky.

4

Explicit Diversification

With the prominence of query ambiguity in web search and the development of advanced query understanding techniques, new diversification approaches taking a query-oriented perspective emerged. As illustrated in Figure 4.1, such approaches explicitly model the aspects underlying a query, notably to directly seek an improved coverage of these aspects among the ranked documents. As a result, modelling aspects that accurately reflect the possible information needs underlying an ambiguous query is a major challenge for such explicit diversification approaches. In this chapter, we describe the most representative of these approaches from the literature. Similarly to Chapter 3, we further organise these approaches as either novelty-based, coverage-based, or hybrid.

4.1 Novelty-based Approaches

Explicit search result diversification approaches have traditionally adopted a coverage-based or hybrid strategy, in order to maximise the retrieved documents' coverage of the explicitly represented query aspects. In contrast, as discussed in §3.1, pure novelty-based strategies were mostly deployed with an implicit aspect representation, with

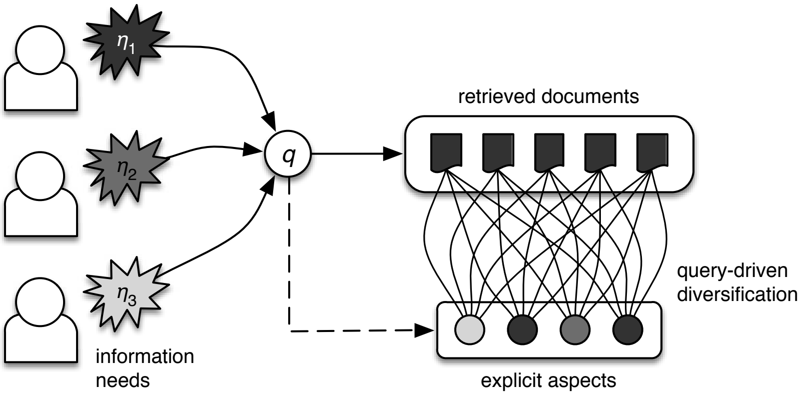

Figure 4.1: Schematic view of explicit diversification approaches.

the aim of promoting documents with dissimilar contents compared to those at higher ranking positions. On the other hand, the prevalence of different aspect representations among the existing approaches precluded a direct comparison between coverage and novelty as diversification strategies. As a result, it was generally unclear whether the differences in diversification performance commonly observed between these approaches was due to their underlying aspect representation (i.e., implicit vs. explicit) or to their adopted diversification strategy (i.e., novelty-based, coverage-based, or hybrid).

To better understand the role played by different aspect representations and different diversification strategies, Santos et al. [2012b] proposed adaptations of existing implicit novelty-based diversification approaches to leverage explicit aspect representations. As a result, they could assess the impact of either dimension in isolation from the other. In particular, given a query q with a set of aspects $\mathcal{S}$, with $|\mathcal{S}| = k$, they explicitly represented each retrieved document $d \in \mathcal{R}_q$ as a k-dimensional vector $\mathbf{d}$ over the aspects $\mathcal{S}$. In particular, the i-th dimension of the vector $\mathbf{d}$ was defined such that:

$$\mathbf{d}_i = f(d, s_i), \tag{4.1}$$

where the function f estimates how well the document d satisfies the aspect $s_i \in \mathcal{S}$. Different measures of the document-aspect association can be used, depending on how the aspects underlying the query are identified, e.g., based on reformulations mined from a query log or on categories derived from a classification taxonomy.

Based upon the explicit aspect representation formalised in Equation 4.1, Santos et al. [2012b] introduced explicit versions of two well-known implicit diversification approaches from the literature, namely, MMR (Equation 3.1) and MVA (Equation 3.4). In particular, the explicit version of MMR, denoted xMMR, was formalised as:

$$f_{\mathrm{xMMR}}(q, d, \mathcal{D}_q) = \lambda \, f_1(q, d) - (1 - \lambda) \max_{\mathbf{d}_j \in \mathcal{D}_q} f_2(\mathbf{d}, \mathbf{d}_j), \tag{4.2}$$

where $f_1(q, d)$ estimates the relevance of d with respect to the query q and $f_2(\mathbf{d}, \mathbf{d}_j)$ estimates the similarity between the explicit representations of d and each of the documents already selected in $\mathcal{D}_q$, with the parameter λ controlling the balance between the two scores.

An explicit version of MVA, denoted xMVA, was devised similarly. In particular, the objective function of xMVA was formalised as:

$$f_{\mathrm{xMVA}}(q, d, \mathcal{D}_q) = \mu_d - b \, w_i \, \sigma_d^2 - 2 \, b \, \sigma_d \sum_{d_j \in \mathcal{D}_q} w_j \, \sigma_{d_j} \, \rho_{\mathbf{d}, \mathbf{d}_j}, \tag{4.3}$$

where μ_d and σ_d^2 are the mean and variance of the relevance estimates for document d, respectively. The summation component estimates the redundancy of d, based upon the correlation $\rho_{\mathbf{d}, \mathbf{d}_j}$ between explicit representations of this document and each document in $\mathcal{D}_q$. The parameter b controls the balance between relevance, variance, and redundancy.

Through an empirical study, Santos et al. [2012b] contrasted xMMR and xMVA to their implicit counterparts, as well as to pure content-based and hybrid diversification approaches using a variety of explicit aspect representations. Compared to MMR and MVA, they observed no consistent improvements when switching to an explicit aspect representation. On the other hand, for a range of explicit representations held fixed, they concluded that novelty, as a diversification strategy, consistently underperforms when compared to coverage. Through a comprehensive simulation, they noted that novelty fails to reward documents that cover multiple query aspects, which could explain the observed differences in performance. On the other hand, they also noted that novelty plays an important role in hybrid approaches, acting as a tie-breaker between documents with similar levels of coverage.

4.2 Coverage-based Approaches

Along with hybrid approaches, which will be described in §4.3, coverage-based approaches are the most prominent explicit diversification approaches in the literature. While the former approaches adhere strictly to the greedy formulation in Algorithm 2.1, the latter ignore dependencies among the retrieved documents in favour of efficiency.

The first explicit coverage-based diversification approach in the literature was introduced by Radlinski and Dumais [2006]. In particular, they proposed to diversify the documents retrieved for a query according to these documents' coverage of multiple reformulations of the query, mined from a query log. To this end, given a query q, they selected the k queries most likely to follow q across multiple sessions in a query log as a set $\mathcal{S}_q$ of query reformulations. In order to select the τ most diverse documents from the ranking $\mathcal{R}_q$, they enforced a uniform coverage of the identified reformulations. According to this uniform coverage (UC) policy, each reformulation $s \in \mathcal{S}_q$ could be represented by at most τ/k documents, which essentially filtered out documents covering already well covered reformulations, such that:

$$f_{\mathrm{UC}}(q, d, \mathcal{D}_q) = \begin{cases} f(q, d) & \text{if } \exists s \in \mathcal{S}_q \mid d \in \mathcal{R}_s \land |\mathcal{R}_s \cap \mathcal{D}_q| < \tau/k, \\ 0 & \text{otherwise,} \end{cases} \tag{4.4}$$

where $\mathcal{R}_s$ is the set of documents that match s. Despite ensuring a uniform coverage of different reformulations, the selected documents are still ranked by their estimated relevance to q, i.e., $f(q, d)$.

In a similar vein, Capannini et al. [2011] proposed to mine query specialisations (i.e., queries with a more specific representation of the user's information need compared to the initial query [Boldi et al., 2009b]) from a query log in order to guide the diversification process. In particular, they selected the τ most diverse documents from $\mathcal{R}_q$ according to each document's weighted proportional coverage (WPC) of the identified specialisations $s \in \mathcal{S}_q$. More precisely, their approach can be formalised into the following objective function:

$$f_{\text{WPC}}(q, d, \mathcal{D}_q) = \begin{cases} f(q, d) & \text{if } \exists s \in \mathcal{S}_q \mid d \in \mathcal{R}_s \wedge |\mathcal{R}_s \cap \mathcal{D}_q| < p(s|q)\,\tau, \\ 0 & \text{otherwise,} \end{cases} \tag{4.5}$$

where $p(s|q)\,\tau$ is the proportion of the final ranking dedicated to documents matching each specialisation $s \in \mathcal{S}_q$, given each specialisation's likelihood $p(s|q)$. For documents matching a not well represented specialisation s, $f(q, d)$ denotes each document's utility, such that $f(q, d) \propto \sum_{s \in \mathcal{S}_q} p(s|q) \sum_{d_j \in \mathcal{R}_s} \frac{1 - f(d, d_j)}{r(d_j, \mathcal{R}_s)}$, where $\mathcal{R}_s$ is a ranking produced for each specialisation s and $f(d, d_j)$ measures the similarity between d and each document $d_j \in \mathcal{R}_s$, ranked at position $r(d_j, \mathcal{R}_s)$.

4.3 Hybrid Approaches

Hybrid approaches based on explicit query aspect representations have also been successfully investigated in the literature. In particular, these approaches try to improve upon pure coverage-based ones, by diminishing the value of covering already well covered query aspects. As a result, excessive redundancy in the final ranking can be avoided.

The first explicit diversification approach in the literature to adopt a hybrid diversification strategy was introduced by Agrawal et al. [2009]. In particular, they proposed to diversify a document ranking in light of a taxonomy $\mathcal{T}$ of query intents, represented as different categories from the Open Directory Project (ODP).[1] Given the classification of both queries and documents in light of this taxonomy, they proposed an intent-aware selection (IA-Select) mechanism, instantiating the objective function in Algorithm 2.1 according to:

$$f_{\text{IA-Select}}(q, d, \mathcal{D}_q) = \sum_{c \in \mathcal{T}} f(c|q, \mathcal{D}_q)\, f(d|q, c), \tag{4.6}$$

where, for each category $c \in \mathcal{T}$, $f(d|q, c)$ denotes the extent to which the document d covers c, while $f(c|q, \mathcal{D}_q)$ denotes the marginal utility of c given the query q and the documents already in $\mathcal{D}_q$. Intuitively, an

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

already well covered category is deemed less useful, which contributes to promoting novel documents and demoting redundant ones.

Dang and Croft [2012] introduced a diversification approach inspired by electoral processes for achieving a proportional representation of the aspects underlying a query in the produced ranking. Different from the coverage-based proportional coverage approaches of Radlinski and Dumais [2006] and Capannini et al. [2011], described in §4.2, Dang and Croft [2012] defined the concept of a proportionality quotient $\zeta(s|q)$ for each query aspect s, according to:

$$\zeta(s|q) \propto \frac{p(s|q)}{\sum_{d_j \in \mathcal{D}_q} p(d_j|s)},$$

where the numerator represents the amount of “votes” received by the aspect s—for instance, based upon its popularity—and the denominator represents the amount of “seats” already taken by s, based upon how much this aspect is already covered by the documents in $\mathcal{D}_q$. The quotient $\zeta(s|q)$ is at the heart of the objective function of their proposed proportionality model (PM-2), which can be defined as:

$$f_{\text{PM-2}}(q, d, \mathcal{D}_q) = \sum_{s \in \mathcal{S}_q} b_s \, \zeta(s|q) \, p(d|s), \tag{4.7}$$

where, for a given parameter λ, $b_s = \lambda$ if aspect s has the highest quotient $\zeta(s|q)$ among all aspects in $\mathcal{S}_q$, or $b_s = 1 - \lambda$ otherwise, and $p(d|s)$ measures the coverage of document d with respect to s.

Santos et al. [2010b,c] proposed to mine query reformulations from a query log as a close representation of the possible information needs underlying an ambiguous query [Santos and Ounis, 2011]. Such an explicit representation of query aspects, which they called sub-queries, formed the basis of their proposed Explicit Query Aspect Diversification (xQuAD) framework. In particular, xQuAD models the concepts of relevance and diversity as a mixture of probabilities, according to:

$$f_{\text{xQuAD}}(q, d, \mathcal{D}_q) = (1 - \lambda) \, p(d|q) + \lambda \, p(d, \bar{\mathcal{D}}_q|q), \tag{4.8}$$

where $p(d|q)$ and $p(d, \bar{\mathcal{D}}_q|q)$ model the probabilities that the document d is relevant and diverse, respectively, with the parameter λ modelling the trade-off between promoting relevance or diversity. The component

$p(d, \bar{\mathcal{D}}_q|q)$ can be interpreted as the probability that d, but none of the already selected documents in $\mathcal{D}_q$, is relevant to q. To account for the multiple information needs possibly underlying q, this probability can be marginalised over the sub-queries $s \in \mathcal{S}_q$, such that:

$$p(d, \bar{\mathcal{D}}_q|q) \approx \sum_{s \in \mathcal{S}_q} p(s|q)\, p(d|q,s) \prod_{d_j \in \mathcal{D}_q} (1 - p(d_j|q,s)),$$

where $p(s|q)$ denotes the importance of sub-query s in light of all sub-queries in $\mathcal{S}_q$ identified for the query q, $p(d|q,s)$ represents the coverage of the document d with respect to s, and $\prod_{d_j \in \mathcal{D}_q}(1 - p(d_j|q,s))$ quantifies the novelty of d, estimated as the probability that none of the documents in $\mathcal{D}_q$ is relevant to the sub-queries covered by d.

4.4 Summary

In this chapter, we have surveyed diversification approaches that rely on an explicit aspect representation. Such approaches build upon recent advances on query understanding, by representing the possible information needs underlying an ambiguous query as multiple query reformulations mined from a query log, or multiple query categories obtained from an existing taxonomy. Explicit approaches have been shown to consistently outperform implicit ones in a standard web search result diversification scenario. In particular, explicit approaches that promote a hybrid of coverage and novelty, such as IA-Select, PM-2, and xQuAD, are regarded as the state-of-the-art in the search result diversification literature. Moreover, these approaches have inspired several advances in the field, as will be introduced in Chapter 6, which will also address applications of diversification in other domains. Meanwhile, in the next chapter, we will describe the currently established methodology, benchmarks, and metrics for evaluating diversity-oriented rankings.

5

Diversity Evaluation

A diverse ranking is one that satisfies the multiple information needs possibly underlying an ambiguous or underspecified query—be these needs from different users or from the same user in different contexts. While traditional search evaluation is challenging, departing from the assumption that a single information need underlies each query arguably renders the evaluation of diversity even more complex. In this chapter, we review the literature on diversity evaluation. In particular, we describe publicly available diversity evaluation benchmarks, produced in the context of the TREC and NTCIR forums, as well as the most prominent metrics used for diversity evaluation.

5.1 Evaluation Benchmarks

Evaluating the effectiveness of ranking approaches is an open challenge. In particular, not only is relevance an ill-understood concept per se [Mizzaro, 1997], but it can also span multiple dimensions [Borlund, 2003], particularly in light of the complex information needs of search users [Broder, 2002, Rose and Levinson, 2004]. Alternative evaluation methodologies have been proposed and tested throughout the years, based upon both implicit and explicit user feedback regarding the relevance of documents ranked in response to a query [Sanderson, 2010].

One of the most established retrieval evaluation methodologies abstracts away from the specificities of individual users, instead relying on the relevance assessment of expert judges to produce an evaluation *benchmark* [Voorhees, 2007]. Such a methodology was pioneered by Cleverdon [1967] at the College of Aeronautics, Cranfield, UK, in their experiments to assess the effectiveness of multiple indexing approaches. While the so-called *Cranfield paradigm* may limit the assessment of relevance in context [Teevan et al., 2007], it dramatically improves the reproducibility of the resulting evaluation, by allowing multiple ranking approaches to be tested on a common benchmark [Voorhees and Harman, 2005]. Moreover, it is estimated that such a methodology has fostered around one third of all improvements in web search ranking from 1999 to 2009 [Rowe et al., 2010].

A benchmark test collection comprises three main components: a corpus of documents, a set of queries, and a set of relevance assessments, which function as a mapping between each query and the documents deemed as relevant for this query [Voorhees, 2007]. While search systems have greatly benefited from the controlled evaluation offered by benchmark test collections, query ambiguity has been largely ignored by early test collections. In practice, the assumption that the user's query represents a single information need reduces the complexity of the underlying evaluation, ensuring that different systems are evaluated with respect to an unambiguously defined information need [Cleverdon, 1991]. However, as pointed out by Spärck-Jones et al. [2007], this assumption does not hold in modern search scenarios, particularly with the high incidence of short and ambiguous queries. Moreover, such queries can negatively impact search effectiveness [Sanderson, 2008].

In order to address such a limitation of the established evaluation paradigm, Spärck-Jones et al. [2007] argued for the development of test collections that explicitly account for queries with different levels of ambiguity. In particular, they claimed that such a test collection should consider each query as representing an ensemble of information needs, as opposed to a single need. In turn, such needs should reflect the interests of the population of users that could have issued the query. Finally, the relevance of each ranked document should be judged separately for each information need, so as to enable the assessment of the effectiveness of the whole ranking at satisfying the multiple needs.

Diversity evaluation is typically operationalised by representing the possible information needs underlying a query as multiple query aspects.[1] Early attempts to build a test collection for diversity evaluation were made at the Text REtrieval Conference (TREC), one of the major forums for research in information retrieval [Voorhees, 2007], which can be seen as a modern instantiation of the Cranfield paradigm. In particular, the TREC 6-8 Interactive tracks [Over, 1997, 1998, Hersh and Over, 1999] investigated a search task called "aspect retrieval", which involved finding documents that covered as many different aspects of a given query as possible. In this evaluation campaign, a total of 20 queries were adapted from the corresponding years of the TREC Ad hoc tracks [Voorhees and Harman, 1997, 1998, 1999]. Together, these queries comprise a total of 398 aspects, identified by TREC assessors, with relevance assessments provided at the aspect level. Figure 5.1 illustrates one of such queries, 353i, along with some of its identified aspects, denoted "sub-topics" in the TREC jargon.

```
<topic number="353i">
  <query> antarctic exploration </query>
  <description>
    Identify systematic explorations and scientific investigations of
    Antarctica, current or planned.
  </description>
  <subtopic number="1"> mining prospection </subtopic>
  <subtopic number="2"> oil resources </subtopic>
  <subtopic number="3"> rhodium exploration </subtopic>
  <subtopic number="4"> ozone hole / upper atmosphere </subtopic>
  <subtopic number="5"> greenhouse effect </subtopic>
    ...
</topic>
```

Figure 5.1: TREC-7 Interactive track, query 353i and its aspects.

[1] Note that the aspect representation adopted by a diversification approach does not necessarily reflect the ground-truth aspect representation used for evaluation.

By relying on expert judges to identify query aspects from the retrieved documents [Lagergren and Over, 1998], the TREC Interactive track test collection arguably lacks in plausibility and completeness in light of the actual information needs of the population of users issuing a query [Radlinski et al., 2010a]. In order to overcome this limitation, Radlinski et al. [2010b] proposed to identify more realistic query aspects for diversity evaluation from the query and click logs of a commercial web search engine. In their approach, candidate query aspects were selected as queries that frequently co-occurred with the initial query across multiple sessions in the query log. Candidate aspects with a low transition probability after a two-step random walk on the bipartite query-document click graph [Craswell and Szummer, 2007] were filtered out. The remaining candidates were then clustered using a graph partitioning algorithm [Blondel et al., 2008]. The highest-scoring aspects from different clusters were shown to better reflect real user needs compared to aspects proposed by expert judges [Radlinski et al., 2010b,a]. As a result, these mined query aspects served as the basis for a new test collection, developed in the context of the TREC 2009-2012 Web tracks [Clarke et al., 2009a, 2010, 2011b, 2012].

The diversity task of the TREC 2009-2012 Web tracks provides one of the largest publicly available test collections for diversity evaluation.[2] In particular, these test collections comprise a total of 198 queries encompassing a total of 579 aspects [Clarke et al., 2009a, 2010, 2011b, 2012]. An example TREC Web track query, along with its identified aspects, is shown in Figure 5.2. In contrast to the short description provided by the TREC Interactive track test collection, the TREC Web track aspects include a natural language description of the information need represented by each aspect. Moreover, each aspect is further classified as either informational ("inf") or navigational ("nav") by TREC assessors, depending on the intent of its underlying need [Broder, 2002].

Another test collection for the evaluation of web search result diversification was introduced as part of the NII Testbeds and Community

[2] In the TREC 2013 Web track [Collins-Thompson et al., 2013], the diversity task was replaced by a risk-sensitive task, aimed at assessing the robustness of ranking approaches to queries with distinct characteristics, including ambiguous ones.

```
<topic number="1">
  <query> obama family tree </query>
  <description>
    Find information on President Barack Obama’s family history,
    including genealogy, origins, places and dates of birth, etc.
  </description>
  <subtopic number="1" type="nav">
    Find the TIME magazine photo essay "Barack Obama’s Family Tree".
  </subtopic>
  <subtopic number="2" type="inf">
    Where did Barack Obama’s parents and grandparents come from?
  </subtopic>
    ...
</topic>
```

Figure 5.2: TREC 2009 Web track, query 1 and its aspects.

for Information access Research (NTCIR) project, a series of evaluation workshops initiated in 1999 with the goal of assessing information retrieval in Asian languages, as well as across different languages. For the NTCIR-9 and NTCIR-10 Intent tasks [Song et al., 2011a, Sakai et al., 2013b], two test collections were developed, aimed at evaluating search result diversification on the Chinese and the Japanese Web.[3] In particular, the Chinese collection comprises 197 queries with a total of 1,532 aspects. For the Japanese collection, another 198 queries were developed, including a total of 1,673 aspects. An example Chinese query (translated to English) is shown in Figure 5.3.

Different from the diversity task of the TREC 2009-2012 Web tracks, the NTCIR-9 and NTCIR-10 Intent tasks include graded (i.e., non-binary) relevance assessments at the aspect level. In addition, as shown in Figure 5.3, the identified aspects are assigned non-uniform probabilities, estimated through assessor agreement, in order to place more emphasis on popular aspects during the evaluation [Sakai and Song, 2012]. Finally, these test collections also enable the assessment of the aspects mined for each query. Statistics of all test collections presented in this section are provided in Table 5.1.

[3] In NTCIR-11, which is ongoing as we write, the Intent task was extended into the iMine task, with an emphasis on large-scale evaluation through crowdsourcing.

```
<topic number="0015">
  <query> mozart </query>
  <subtopic number="1" probability="0.241379310344828">
    mozart’s music download
  </subtopic>
  <subtopic number="2" probability="0.241379310344828">
    mozart’s biography
  </subtopic>
  <subtopic number="3" probability="0.241379310344828">
    works by mozart
  </subtopic>
  ...
</topic>
```

Figure 5.3: NTCIR-9 Intent task (Chinese), query 0015 and its aspects.

Table 5.1: Statistics of publicly available test collections for diversity evaluation. The TREC Interactive track (IN6, IN7, IN8) and Web track (WT09, WT10, WT11, WT12) include English queries and documents, whereas the NTCIR Intent task (IT1, IT2) includes both Chinese (ZH) and Japanese (JA) queries and documents.

	TREC Interactive track [Over, 1997, 1998, Hersh and Over, 1999]					
	IN6	IN7	IN8	Total		
# queries	6	8	6	20		
# aspects	84	140	174	398		
# relevants	161	297	352	810		
	TREC Web track [Clarke et al., 2009a, 2010, 2011b, 2012]					
	WT09	WT10	WT11	WT12	Total	
# queries	50	48	50	50	198	
# aspects	199	29	164	187	579	
# relevants	4,941	6,552	5,026	5,559	22,078	
	NTCIR Intent task [Song et al., 2011a, Sakai et al., 2013b]					
	IT1-ZH	IT1-JA	IT2-ZH	IT2-JA	Total-ZH	Total-JA
# queries	100	100	97	95	197	198
# aspects	917	1091	615	582	1,532	1,673
# relevants	23,571	19,841	9,295	5,085	32,866	24,926

5.2 Evaluation Frameworks

Several metrics have been proposed in recent years to evaluate the diversification effectiveness of a document ranking. Given a query q and a cutoff κ, a diversity evaluation metric quantifies the extent to which the top κ documents in a ranking $\mathcal{R}_q$ cover the aspects $\mathcal{A}_q$, representing the information needs $\mathcal{N}_q$ underlying q.

The most straightforward metric for diversity evaluation is perhaps sub-topic recall (SR) [Zhai et al., 2003]. This metric quantifies the amount of unique aspects of the query q that are covered by the top κ ranked documents $d \in \mathcal{R}_q^{(\kappa)}$, according to:

$$\mathrm{SR}(q, \kappa) = \frac{\cup_{d \in \mathcal{R}_q^{(\kappa)}} |\mathcal{A}_q \cap \mathcal{A}_d|}{|\mathcal{A}_q|}, \tag{5.1}$$

where $\mathcal{A}_q$ is the set of relevant query aspects and $\mathcal{A}_d$ is the set of aspects for which the retrieved document $d \in \mathcal{R}_q^{(\kappa)}$ is relevant.

A limitation of sub-topic recall is that it does not take into account the probability of different aspects given the query. Ideally, this probability should reflect the fraction of the population that is interested in the information need represented by each aspect. Alternative evaluation frameworks that account for the (potentially non-uniform) probability of different aspects have been proposed in the literature. In common, these frameworks generate diversity-oriented metrics as a natural extension of relevance-oriented metrics in the presence of multiple query aspects. One such framework was proposed by Agrawal et al. [2009]. In particular, they defined an "intent-aware" (IA)[4] diversity evaluation metric Eval-IA(q, κ) as the *expected value* of its counterpart relevance-oriented metric Eval(a, κ), with $a \in \mathcal{A}_q$, such that:

$$\text{Eval-IA}(q, \kappa) = \sum_{a \in \mathcal{A}_q} p(a|q) \mathrm{Eval}(a, \kappa), \tag{5.2}$$

where $p(a|q)$ is the probability of observing the aspect a given the query q, and Eval(a, κ) assumes that a is the only relevant aspect of q.

[4] Agrawal et al. [2009] use "intent" in the sense of "information need". In this survey, we adopt the traditional definition of "intent" as a *property* of an information need (e.g., informational, navigational), in the sense proposed by Broder [2002], and instead refer to the information needs underlying a query as "aspects".

A limitation of intent-aware metrics is that they do not enforce a high coverage of multiple query aspects by design. As a result, these metrics may completely ignore aspects with a low probability $p(a|q)$. In the extreme case, they may end up maximally rewarding a ranking that covers only a single yet dominant aspect. An option to enforce the coverage of multiple query aspects within the intent-aware framework is to compute the expected value of a *cascade* metric [Clarke et al., 2011a]. Cascade metrics penalise redundancy by modelling the behaviour of a user who stops inspecting the ranking once a relevant document is observed [Craswell et al., 2008]. As an indirect result, these metrics encourage the coverage of non-redundant aspects. One such metric is expected reciprocal rank (ERR) [Chapelle et al., 2009], defined as:

$$\mathrm{ERR}(q, \kappa) = \sum_{i=1}^{\kappa} \frac{1}{i} \prod_{j=1}^{i-1} (1 - p_j)\, p_i, \tag{5.3}$$

where p_i denotes the probability that the i-th document is relevant to the query, in which case $\prod_{j=1}^{i-1}(1-p_j)$ denotes the probability that none of the documents ranked higher than the i-th document is relevant. In practice, p_i is defined as a function of the relevance grade g_i of the i-th document, i.e., $p_i = (2^{g_i} - 1)/2^{g_{\max}-1}$, where $g_{\max}$ is the maximum grade considered. This metric can be extended into its intent-aware counterpart, ERR-IA [Chapelle et al., 2011b], according to:

$$\text{ERR-IA}(q, \kappa) = \sum_{a \in \mathcal{A}_q} p(a|q) \mathrm{ERR}(a, \kappa), \tag{5.4}$$

where $\mathrm{ERR}(a, \kappa)$ is computed separately for each aspect $a \in \mathcal{A}_q$, under the assumption that none of the other query aspects is of interest.

An alternative evaluation framework that enforces a high coverage of multiple query aspects by design was proposed by Clarke et al. [2008]. In particular, instead of computing the expected *value* of a relevance metric across each of the multiple aspects $\mathcal{A}_q$ underlying the query q, as in Equation 5.2, they proposed to extend this metric to leverage the expected *gain* over multiple aspects, which they called "information nuggets". The introduced family of diversity metrics, here denoted "N" metrics, can be formalised according to:

$$\text{N-Eval}(q, \kappa) = \mathrm{Eval}(\mathcal{A}_q, \kappa), \tag{5.5}$$

where $\mathrm{Eval}(\mathcal{A}_q, \kappa)$ denotes a traditional graded relevance metric, with the gain of the i-th document computed by aggregating the aspect-specific gains $g_{i|a}$, according to $g_i = \sum_{a \in \mathcal{A}_q} p(a|q)\, g_{i|a} (1-\alpha)^{\sum_{j=1}^{i-1} g_{j|a}}$. The factor $(1-\alpha)^{\sum_{j=1}^{i-1} g_{j|a}}$ penalises redundancy by diminishing the value of covering the aspect a, according to how much this aspect is already covered by the documents ranked ahead of the i-th document. In practice, an advantage of this framework over the intent-aware framework of Agrawal et al. [2009] is that the metric $\mathrm{Eval}(\mathcal{A}_q, \kappa)$ is computed for a single ranking rather than for multiple separate rankings.

Clarke et al. [2008] instantiated the nugget-based framework to produce the well-known α-DCG metric [Clarke et al., 2008], which extends the traditional DCG metric [Järvelin and Kekäläinen, 2002] by using aspect-specific binary gains $g_{i|a}$, according to:

$$\alpha\text{-DCG}(q, \kappa) = \sum_{i=1}^{\kappa} \frac{\sum_{a \in \mathcal{A}_q} g_{i|a} (1-\alpha)^{\sum_{j=1}^{i-1} g_{j|a}}}{\log_2(i+1)}, \tag{5.6}$$

where the parameter $\alpha \in [0, 1)$ controls the amount of penalisation of the gain $g_{i|a}$. In particular, a value of $\alpha \to 1$ results in the maximum penalisation, whereas $\alpha = 0$ reduces to the standard DCG, with the number of covered aspects $\sum_{a \in \mathcal{A}_q} g_{i|a}$ used as the gain at rank i.

Sakai et al. [2010] argued that the explicit penalisation of redundancy performed by nugget-based metrics such as α-DCG may be undesirable, given that relevance metrics such as DCG already have a rank-based discount aimed at this purpose [Järvelin and Kekäläinen, 2002]—the log-based discount in the summand of Equation 5.6. To avoid such a double discount, they proposed "D" metrics, by aggregating non-discounted, non-binary gains $g_{i|a}$ in order to express the degree of relevance of the i-th document with respect to aspect a. To further emphasise the coverage of aspects with a low probability in the ranking, they further proposed to linearly interpolate a D metric with sub-topic recall, defined in Equation 5.1. The resulting metric, which they called a "D♯" metric, can be defined as:

$$\begin{aligned} \text{D}\sharp\text{-Eval}(q, \kappa) = {} & \gamma\, \mathrm{SR}(q, \kappa) \\ & + (1-\gamma)\text{D-Eval}(q, \kappa), \end{aligned} \tag{5.7}$$

where the parameter γ controls the balance between $\mathrm{SR}(q, \kappa)$ and D-Eval(q, κ). Typically, this parameter is set as $\gamma = 0.5$, as it was shown to have little impact in D♯-Eval(q, κ), primarily because $\mathrm{SR}(q, \kappa)$ and D-Eval(q, κ) are highly correlated with each other [Sakai et al., 2010].

Extended metrics have also been proposed in recent years, accounting for dimensions of the diversification problem not addressed by the metrics described thus far. For instance, Clarke et al. [2009b] proposed a metric that explicitly distinguishes between aspects related to different interpretations of the user's query. Their basic intuition was that, while a user may be interested in multiple aspects of a given interpretation, only one such interpretation should be of interest. To exploit this intuition, they built upon the rank-biased precision (RBP) metric [Moffat and Zobel, 2008], a graded relevance metric with a parameter β denoting the (fixed) probability that a user will inspect a further document. The higher the value of β, the more persistent the user. In particular, RBP was extended with a discount factor that penalises redundancy, similarly to α-DCG [Clarke et al., 2008]. The resulting metric, novelty- and rank-biased precision (NRBP), was defined as:

$$\mathrm{NRBP}(q, \kappa) = \frac{(1-(1-\alpha)\beta)}{\beta} \sum_{i=1}^{\kappa} \beta^i \sum_{\varphi \in \Omega_q} \frac{p(\varphi|q)}{|\mathcal{A}_\varphi|} \sum_{a \in \mathcal{A}_\varphi} g_{i|a} (1-\alpha)^{\sum_{j=1}^{i-1} g_{j|a}}, \tag{5.8}$$

where Ω_q is the set of possible interpretations of the query q, and $\mathcal{A}_\varphi$ is the set of aspects associated with each interpretation $\varphi \in \Omega_q$, in which case $g_{i|a}$ denotes the (binary) relevance grade of the i-th document with respect to the aspect $a \in \mathcal{A}_\varphi$. Interpretations follow a non-uniform distribution $p(\varphi|q)$, whereas the distribution of aspects for a given interpretation is assumed to be uniform. Analogously to α-DCG in Equation 5.6, $(1-\alpha)^{\sum_{j=1}^{i-1} g_{j|a}}$ penalises the coverage of already well covered interpretation-aspect pairs, with the parameter α controlling the amount of penalisation. The extra parameter β models users with different patience levels, similarly to the standard RBP metric.

Sakai [2012] proposed to extend the IA and D frameworks in order to take into account the *intent* of different query aspects. In particular, for the extended D framework, he computed the gain at rank i by distinguishing between informational and navigational aspects, according to $g_i = \sum_{a \in \mathcal{A}_q} p(a|q)\, g_{i|a}(1 - \mathbf{1}_{\mathcal{A}_q^{\text{nav}}}(a)\, \mathbf{1}_{\cup_{j=1}^{i-1} \mathcal{A}_{d_j}}(a))$, where the indicator functions $\mathbf{1}_{\mathcal{A}_q^{\text{nav}}}(a)$ and $\mathbf{1}_{\cup_{j=1}^{i-1} \mathcal{A}_{d_j}}(a)$ denote whether the aspect $a \in \mathcal{A}_q$ is navigational and whether it is covered by any document ranked ahead of the i-th document. His assumption was that redundancy should be penalised for navigational aspects, but not for informational ones. An analogous extension was proposed for the IA framework, by interpolating the expected value of informational- and navigational-oriented metrics over the corresponding subsets of query aspects.

Chandar and Carterette [2013] noted that performing relevance assessments at the query aspect level may be costly, given the need to assess each retrieved document with respect to each of the multiple aspects underlying a query. Instead, they proposed an alternative framework that exploits multiple assessors' preferences for different documents retrieved for a query. In particular, for each assessor, they obtained pairwise document preferences conditioned on the documents previously observed by the assessor. Assuming that each assessor has a potentially different information need in mind for the query, considering preferences manifested by multiple assessors conveys each document's coverage of distinct query aspects. On the other hand, conditioning these preferences on the previously observed documents further conveys the document's novelty with respect to the previous ones. Through an empirical analysis, expected utility metrics exploiting these intuitions were shown to correlate well with current evaluation metrics that leverage explicit relevance assessments at the aspect level.

5.3 Meta Evaluation

In addition to developing diversity metrics, much effort has been invested in validating such metrics. For instance, Clarke et al. [2011a] analysed the *discriminative power* of diversity metrics, a property that reflects the extent to which a metric can distinguish between pairs of rankings [Sakai, 2006]. Using the runs submitted to the TREC 2009 Web track [Clarke et al., 2009a], they observed that sub-topic recall (Equation 5.1) has the highest discriminative power compared to the other considered diversity metrics. Intent-aware and cascade metrics, on the other hand, showed a discriminative power inferior to that observed for average precision [Harman, 1993], a relevance-oriented metric. Relatedly, Leelanupab et al. [2012] observed that the discriminative power of α-DCG can be improved by appropriately setting the parameter α on a per-query basis to ensure that the coverage of novel aspects is rewarded higher than the coverage of redundant ones.

Golbus et al. [2013] observed that the diversification effectiveness of a system for a query is influenced by the difficulty of this query, quantified based on the number of relevant aspects underlying the query, and the distribution of relevant documents per aspect. While diversity difficulty impacts the diversification effectiveness that can be attained by any system, they also noted that this effectiveness is highly dependent on the system's ad hoc effectiveness. Accordingly, they proposed to assess the *sensitivity* of a diversity evaluation metric as the ability of the metric to distinguish between systems with similar ad hoc effectiveness. In a simulation with rankings of perfect ad hoc effectiveness, they showed that existing metrics have low sensitivity, which could be improved by explicitly accounting for the notion of diversity difficulty.

Ashkan and Clarke [2011] analysed the *informativeness* of diversity metrics, which reflects the extent to which a metric predicts the actual distribution of relevant documents. Using the maximum entropy method to estimate the most plausible relevance distribution according to a given metric [Aslam et al., 2005], they found that intent-aware cascade metrics (which reward coverage and novelty) are more informative than their pure cascade counterpart (which only rewards novelty), with ERR-IA [Chapelle et al., 2011b], described in Equation 5.4, showing the highest informativeness among all considered metrics.

Sanderson et al. [2010] investigated the *predictive power* of diversity metrics, based on how well these metrics correlate with the behaviour of actual users. In their study, 296 subjects were hired through crowdsourcing to express their preference between pairs of runs submitted to the TREC 2009 Web track [Clarke et al., 2009a]. The runs in each pair were also evaluated according to multiple diversity metrics. Their analysis showed a high agreement between the prediction of several diversity metrics and the users' preferences, with no significant difference in predictive power between the considered metrics.

Carterette [2009] analysed the *optimality* of the normalisation component of cascade metrics. In particular, producing an ideal ranking for normalising such metrics is an NP-hard problem, as discussed in §2.5. Since the ideal ranking is typically computed using the greedy approximation in Algorithm 2.1, a natural question is whether the produced evaluation scores are affected by a sub-optimal normalisation. Fortunately, an analysis of real and simulated topic sets and aspect relevance assessments showed that the greedy and optimal evaluation normalisations agree in 93% and 85% of the cases, respectively.

Sakai [2013] analysed the *reusability* of the currently available test collections for search result diversification. In particular, despite their large sizes, these collections are built from relatively shallow pools, comprising the union of the top 20 to 40 documents retrieved by each contributing ranking. Through a leave-one-out simulation [Zobel, 1998], he observed that rankings not pooled during the construction of these test collections tend to retrieve a substantial amount of documents that were not judged, and hence have their diversity effectiveness underestimated. As an alternative, he noted that condensed-list versions of existing evaluation metrics [Sakai, 2007], which discard unjudged documents, are less affected when reusing a test collection.

5.4 Summary

In this chapter, we have detailed the currently established methodology for evaluating diversity-oriented rankings. In particular, we described publicly available benchmark test collections for diversity evaluation in web and newswire search. In addition, we introduced different evaluation frameworks and the most used evaluation metrics based on them. Lastly, we discussed quality properties of these metrics, including their discriminative power, sensitivity, informativeness, predictive power, optimality, and reusability. In the next chapter, we will present advanced approaches recently proposed to search result diversification, as well as applications of diversification beyond web search.

6

Advanced Topics and Applications

Several approaches have been introduced in recent years to tackle the search result diversification problem. As discussed in the previous chapters, these approaches employ a variety of strategies for diversifying the search results, by promoting documents with high novelty and coverage of multiple query aspects, identified either implicitly or explicitly. In this chapter, we discuss advanced topics that are relevant to several of these approaches, including the prediction of query ambiguity and the identification of query aspects. In addition, we discuss applications for diversification in domains other than web search, as well as the need for diversifying the search results across multiple search verticals.

6.1 Query Ambiguity Detection

Query ambiguity is generally acknowledged as one of the reasons for poorly performing retrieval systems [Buckley, 2004]. Indeed, as an inherently limited representation of a more complex information need, every query can be arguably considered ambiguous to some extent [Cronen-Townsend and Croft, 2002]. Nevertheless, to assess the extent to which ambiguity is present in web search queries, Song et al.

[2009] conducted a user study based upon 60 queries sampled from the log of a commercial search engine from August 2006. In their study, five assessors manually classified each of these queries as either ambiguous, underspecified, or clear. While a high assessor agreement (90%) was observed for judging whether a given query was ambiguous or not, distinguishing between underspecified and clear queries turned out to be substantially more difficult. Nonetheless, based on the demonstrated feasibility of the former case, they proposed a binary classification approach to automatically identify ambiguous queries. Based on the learned classification model, they estimated that 16% of the queries in their entire query log sample were ambiguous. Surprisingly consensual figures were obtained by assessing the click entropy of web search queries [Clough et al., 2009], and by matching such queries against Wikipedia disambiguation pages [Sanderson, 2008].

Accurately predicting the level of ambiguity of a query can help inform improved diversification strategies. In particular, different levels of ambiguity may entail different needs for diversifying the search results. For instance, the query *"bond"* could have different interpretations, including the financial instrument for debt security, the classical crossover string quartet "Bond", or Ian Fleming's secret agent character "James Bond". In the absence of any knowledge about the user's context or preferences, a more aggressive diversification could improve the chance of returning at least one document relevant to each of these interpretations. In contrast, the query *"james bond"* is arguably less ambiguous, since its interpretation is clear, even if the exact information the user is looking for is not. For example, the user might be interested in several, not necessarily mutually exclusive, aspects such as the history of the character, reviews for the last film, etc. For such a query, a more lenient diversification could suffice. At the extreme, a query like *"james bond spectre website"* has an arguably clear meaning, i.e., the website of the latest film in the series, which may well be addressed by a standard relevance-oriented ranking approach.

Departing from the assumption that all queries should be equally amenable to diversification, Santos et al. [2010a] investigated the optimality of the balance between promoting relevance or diversity across different queries. In particular, they observed that the optimal balance drastically varied from query to query, indeed confirming that different queries could benefit from more of less aggressive diversification strategies. To exploit this intuition, they proposed to learn an effective setting for the diversification trade-off parameter λ, which underlies several diversification approaches in the literature, such as MMR (Equation 3.1) and xQuAD (Equation 4.8), on a per-query basis. To predict an effective trade-off λ_q for a given query q, they employed a lazy regression approach based on nearest neighbours, such that:

$$\lambda_q = \frac{1}{k} \sum_{q_i \in \Gamma_q^k} \lambda_{q_i}^*, \tag{6.1}$$

where Γ_q^k comprises the k nearest neighbouring training queries to q, and $\lambda_{q_i}^*$ denotes the (known) optimal value of λ for the query q_i. In order to identify neighbouring queries, a variety of query features was employed, inspired by tasks such as query log mining and query performance prediction. Experiments using benchmark test collections from the diversity task of the TREC Web track, described in §5.1, demonstrated the effectiveness of this selective approach. In particular, significant improvements were observed for both MMR and xQuAD when leveraging per-query trade-offs instead of a uniform setting of the trade-off for all queries. Moreover, additional experiments using an oracle predictor demonstrated further room for improvements.

6.2 Query Aspect Mining

One of the pillars of every search result diversification approach is the ability to identify query aspects that reflect real users' information needs. In this chapter, we overview several query aspect mining approaches that mine external resources, such as query logs, as well as approaches that seek to identify plausible aspects from alternative sources, such as the initially retrieved documents themselves. In addition, we describe approaches dedicated to diversifying the identified aspects, as well as to predicting the intent underlying each aspect.

6.2.1 Aspect Mining from Query Logs

Web search queries are typically short, ill-defined representations of more complex information needs [Jansen et al., 1998]. As a result, they can lead to unsatisfactory retrieval performance. Query suggestions have been introduced as a mechanism to alleviate this problem. Such a mechanism builds upon the vast amount of querying behaviour recorded by search engines in the form of query logs, in order to suggest related queries previously issued by other users with a similar information need [Silvestri, 2010]. Indeed, a study by Kruschwitz et al. [2013] on website search found that methods of query suggestions based on query logs significantly outperform methods based on obtaining suggestions from the contents of documents within a target corpus.[1] The mined suggestions can be exploited in a variety of ways. For instance, a suggestion identified with high confidence can be considered for automatically rewriting the user's initial query [Jones et al., 2006]. Alternatively, a few high quality suggestions can be offered to the user as alternatives to the initial query [Baeza-Yates et al., 2004], or to help diversify the documents originally retrieved for this query.

Several approaches have been recently proposed to infer the importance of a candidate suggestion for a given query based on these queries' textual similarity, their co-occurrence in common sessions, or their common clicked URLs [Silvestri, 2010]. For instance, Jones et al. [2006] proposed to generate candidate suggestions from co-session queries with a common substring. The strength of the relationship between the query and each suggestion was estimated by leveraging various similarity features, such as the edit distance and the mutual information between these queries. Analogously, Wang and Zhai [2008] proposed to mine term association patterns from a query log. Their approach analysed the co-occurrence of terms in multi-word co-session queries and built a translation model in order to mine query suggestions.

A session-based approach was proposed by Fonseca et al. [2003] for mining query suggestions. In particular, they deployed an association rule mining algorithm in order to identify query pairs with sufficient

[1] We note that they did not experiment with query suggestions derived from anchor text, which has shown promising results, as we will discuss in §6.2.2.

co-occurrence across multiple sessions. Such association rules were then used as the basis for identifying query suggestions from a query log. Relatedly, Zhang and Nasraoui [2006] exploited the sequence of queries in a query log session. In particular, their approach created a directed graph with queries as nodes, and with edges connecting consecutive queries in each session, weighted by these queries' textual similarity. A candidate suggestion for a given query was then scored based on the length of the path between the two queries, accumulated across all sessions where the query and the suggestion co-occurred.

An early work on mining query suggestions from query logs can be attributed to Beeferman and Berger [2000], who performed clustering in a click log in order to identify queries related to the user's entered query. Another click-based approach was proposed by Baeza-Yates et al. [2004]. In particular, they proposed to cluster queries represented using the terms present in the URLs clicked for these queries. Given an input query, candidate suggestions from the same cluster as the query were then weighted based on their similarity to the query and their success rate, as measured by their fraction of clicked documents in a query log. Relatedly, Mei et al. [2008] exploited random walks on a bipartite query-click graph. To this end, they weighted a candidate suggestion for a query based on its "hitting" time (i.e., the time it took for the node representing this query suggestion to be visited for the first time) for a random walk starting from the input query. Similarly, Boldi et al. [2009a] proposed to weight candidate suggestions by performing a short random walk on different slices of a query-flow graph, a query transition graph with edges classified as generalisations, specialisations, error corrections, or parallel moves [Boldi et al., 2008]. In terms of deriving useful suggestions, specialisations were in general found to be the most effective.

Clustering is an ongoing trend in query understanding in general, and in the identification of query suggestions in particular. For instance, Radlinski et al. [2010b] used clusters of queries that lead to clicks on similar sets of documents to help assessors explain the information need behind those queries. This clustering method was also used to generate the queries and the corresponding ground-truth aspects used for evaluating diversity-oriented ranking approaches in the TREC Web track, as discussed in §5.1. Similarly, Dang et al. [2011] demonstrated the use of the agglomerative clustering of queries for identifying aspects underlying an effective diversification. We also note the work of Wu et al. [2011], whose ASPECTOR system combines vertical category information and clustering techniques for aspect identification, supplemented with the use of structured "infobox" data from Wikipedia to gather aspects for rare queries. This was shown to be effective compared to the related searches proposed by Google and Bing.

Overall, random walk approaches are generally regarded as the state-of-the-art in the literature dedicated to the query suggestion problem [Silvestri, 2010]. Despite their relative success, most of these approaches share a common shortcoming, namely, they underperform and can even fail to produce any relevant suggestion for queries with sparse or no past usage in a query log, which amount to a substantial fraction of the web search traffic [Downey et al., 2007]. In order to overcome this issue, Szpektor et al. [2011] proposed the notion of query template, a generalisation of a query in which entities are replaced with their type. By enriching the query-flow graph [Boldi et al., 2008] with query templates, their approach was able to effectively generate suggestions for long-tail queries. A different approach aimed at tackling query sparsity was proposed by Broccolo et al. [2012]. In particular, they proposed to index each query in a query log as a *virtual document* comprising the terms in the query itself and those of other queries from common sessions. As a result, they cast the query suggestion problem as an efficient search over the inverted index of virtual documents.

Santos et al. [2013] built upon the query representation strategy proposed by Broccolo et al. [2012], which was shown to perform at least as effectively as the state-of-the-art query-flow graph approach of Boldi et al. [2009a] for head queries, while consistently outperforming it for queries with little or no past evidence [Broccolo et al., 2012, §4.4]. Inspired by this approach, Santos et al. [2013] devised a *structured* virtual document representation, by treating terms from different sources as distinct *fields*. Besides the candidate suggestion itself and its co-session queries, they also leveraged evidence from queries that shared at least one click with the suggestion. This enriched representation provided multiple criteria for ranking suggestions with respect to a query, which they encoded as query-dependent features in a unified ranking model automatically learned from training data. To further improve this model, they proposed several query-independent features as quality indicators for a candidate suggestion. Their proposed approach was evaluated with respect to the usefulness of the produced suggestions for ad hoc search as well as for search result diversification, when used as sub-queries within the xQuAD framework (Equation 4.8). In particular, they found that it was effective at identifying suggestions even for completely unseen queries in the query log.

6.2.2 Aspect Mining from Alternative Sources

As discussed in the previous sections, a considerable number of queries is completely new to a search engine, having no past history in the search engine's query logs [Downey et al., 2007]. As a result, alternative approaches should be considered to mine query aspects from other sources. For instance, Bhatia et al. [2011] proposed an approach for deriving query suggestions without query logs, based on the extraction of phrases from the target corpus. In particular, their goal was to automatically complete a user's query as it was typed, based upon mined phrases with high frequency. Instead of using the contents of the documents, Kraft and Zien [2004] found that the anchor text of the incoming hyperlinks of documents could be used to generated refined queries that could be suggested as quick links to the users. Such an approach built upon the fact that anchor text and queries share similar characteristics [Eiron and McCurley, 2003], probably because they both represent short user-generated textual descriptions of content with a specific information need in mind. Later, Dang and Croft [2010] reached the same conclusion that anchor text could be used as a replacement of query logs for generating effective suggestions.

Rather than relying on documents or anchor text from an entire corpus, the aspects underlying a query might be obtained from the contents of the top ranked documents for this query, in a typical implicit diversification fashion, as discussed in Chapter 3. For instance, Wang et al. [2013] proposed a technique that mines the text surrounding occurrences of query terms to derive aspects. In a similar manner, Vargas et al. [2013] showed how xQuAD could be instantiated for deriving diverse query terms for query expansion, which resulted in improved diversification effectiveness. Finally, we also note the work of Dang and Croft [2013], who proposed to mine individual terms—as opposed to well-formed "query-like" compounds—as candidate aspects for diversification. While their approach demonstrated the effectiveness of directly using individual terms for diversification, the potential benefits of further exploiting term dependences are unclear.

Structured data have also been successfully leveraged as a source of evidence for mining query aspects for diversification. For instance, Bouchoucha et al. [2013] used ConceptNet [Liu and Singh, 2004] to refine the query in order to include related yet diverse concepts. Going further, Zheng et al. [2014] proposed that structured data could be used to better drive the aspect identification process. Their experiments using data from relational databases within an organisation considered ambiguous topics submitted to the organisation's enterprise search engine, and showed that diversification performance was enhanced when the structured data was used. Similar results for web search were also observed when mixing query logs and ODP categories.

Multiple sources have also been mixed together to provide improved evidence for query aspect mining, under the assumption that each data source provides complementary evidence of the user's information need. For instance, He et al. [2012] combined click logs, anchor text, and web n-grams as sources of aspects within a random walk algorithm. Similarly, in an approach which the authors called multi-dimensional diversification, Dou et al. [2011] combined anchor text, query logs, search result clusters and source websites as sources of aspects, obtaining improvements in diversification effectiveness. In particular, for both approaches, the effectiveness of anchor text follows from its usefulness for deriving query suggestions, as discussed above.

6.2.3 Aspect Mining Extensions

Once the different aspects underlying an ambiguous query have been identified, extended analyses can be performed. In this section, we discuss two such analyses, aimed to improve the choice of aspects and their subsequent use by a search result diversification approach.

Regarding the choice of aspects, the better they reflect the multiple possible information needs underlying an ambiguous query, the more useful they are for search result diversification. To this end, diversification approaches can also be used to diversify the selected aspects themselves. For instance, Song et al. [2011b] proposed an approach for diversifying query suggestions based on the similarity of the result lists retrieved by a search engine for different suggestions. Relatedly, Kharitonov et al. [2013] noted that the amount of diversification of suggestions that is appropriate for a user's query differed according to the knowledge about the user's context. For instance, for a query with suffix *"bond"*, without any prior knowledge of the user, completion suggestions about *"james bond"* and *"financial bond"* would both be appropriate. However, *"financial bond"* would be inappropriate if the user's previous query was *"bond films"* or *"bond actors"*. In order to contextualise and diversify query suggestions, they introduced a framework that takes into account both recent query history as well as the documents clicked and skipped by the user as contextual evidence. Results based on a sample query log from the Yandex search engine attested the effectiveness of their proposed approach. In a similar vein, Kim and Croft [2014] showed how xQuAD (Equation 4.8) could be instantiated to identify diverse query suggestions, based on the documents already retrieved for each suggestion. Indeed, if a query suggestion is novel, then it will bring documents not brought by query suggestions that have already been selected. In terms of coverage, if a query suggestion is related to the original query, it should cover documents similar to those covered by the original query.

Regarding the use of mined aspects for diversification, a further extension relates to predicting the intent of each aspect. In particular, Broder [2002] proposed a taxonomy of information needs in web search, later extended by Rose and Levinson [2004], categorising the intent underlying each information need as either navigational, informational, or transactional. Such intents have been previously shown to affect the estimation of relevance [e.g., Kang and Kim, 2003, Geng et al., 2008, Peng et al., 2010]. Building upon this intuition, Santos et al. [2011b] proposed an intent-aware diversification approach, aimed to exploit the intents underlying different aspects in order to improve existing explicit diversification approaches. For instance, while *"james bond actors"* could be arguably considered an informational aspect, *"james bond spectre website"* has a navigational nature. To exploit this intuition, Santos et al. [2011b] developed intent classifiers using a range of aspect features, and learned ranking models specifically targeted to each individual intent. As a result, they could estimate the coverage and novelty of documents with respect to a given aspect by leveraging the learned model most suitable to the predicted intent of this aspect. Their approach was shown to improve the diversification effectiveness of both IA-Select (Equation 4.6) and xQuAD (Equation 4.8).

6.3 Diversity across Verticals

Thus far, we have discussed the impact of ambiguity and redundancy mostly in the context of web search. Nevertheless, other search verticals (e.g., news, image, product search engines as well as recommender systems) can also suffer from these problems. As a result, such verticals form natural application scenarios for search result diversification.

Regarding diversification in different verticals, several of the implicit diversification approaches introduced in Chapter 3 were proposed in the context of newswire search [e.g., Carbonell and Goldstein, 1998, Zhai et al., 2003, Wang and Zhu, 2009, Chen and Karger, 2006]. In the context of image search, van Leuken et al. [2009] proposed to cluster the retrieved images using visual features. Their intuition was that representative images from distinct clusters could form a diverse ranking. A similar approach was proposed by Deselaers et al. [2009], however mixing both textual and visual features. In a different vein, Paramita et al. [2009] proposed to diversify image search results spatially, by leveraging location information associated to every image.

Structured search verticals, such as document [Zheng et al., 2011a] and expert search [Plachouras, 2011] in an enterprise setting, have also benefited from search result diversification. As another example, Vee et al. [2008] proposed to diversify a ranking of products satisfying query predicates, by devising tree-traversal algorithms to efficiently compare product pairs according to their attributes. Similarly, Gollapudi and Sharma [2009] deployed facility dispersion algorithms to promote diversity, by comparing the products for a query in terms of their categorical distance in a given taxonomy. Other verticals that have benefited from diversification include biomedical search [Yin et al., 2010, Limsopatham et al., 2014], blog search [Demartini, 2011, Santos et al., 2012a], reviews search [Krestel and Dokoohaki, 2011], and film search on IMDB [Demidova et al., 2010].

Recommender systems is another domain that have benefited from diversification, particularly to avoid the over-specialisation of the recommended items [Yu et al., 2009]. In such systems, the user's need is typically represented by an often ambiguous profile of past preferences. Furthermore, redundant or obvious recommendations are of little use, and play against the central goal of recommender systems of promoting the discovery of new items [Jannach et al., 2010]. With this in mind, a few approaches have been proposed to promote diversity in recommender systems. For instance, Ziegler et al. [2005] proposed to promote diversity by iteratively selecting items with minimum similarity to the already selected ones, similarly to the implicit diversification approaches introduced in §3.1. An alternative approach was proposed by Zhou et al. [2010], by favouring rare items in the final recommendation. Lathia et al. [2010] considered diversity from a temporal perspective, by promoting novel items with respect to those recommended in the past. Explicit diversification approaches were also considered by Vargas et al. [2011] and Belém et al. [2013], who adapted IA-Select and xQuAD for recommending films and tags, respectively. Finally, a parallel between evaluation approaches for search and recommendation was drawn by Vargas and Castells [2011]. In particular, they introduced a unified framework as a basis for generating evaluation metrics suitable for assessing diversity in both scenarios.

Modern web search engines include results from multiple search verticals, such as maps, news, images, videos, and products, in addition to the standard retrieval of web pages. Such a unified search approach is generally referred to as universal search or aggregated search [Murdock and Lalmas, 2008, Diaz et al., 2010]. To assess the impact of query ambiguity in aggregated search engines, Santos et al. [2011a] analysed searching behaviour data from four Google verticals, namely, web, image, news, and product search. This analysis showed that the ambiguity of a single query varies considerably across different verticals, both in terms of the aspects underlying the query in different verticals, as well as in terms of the likelihood of these aspects. For instance, *"james bond"* and *"financial bonds"* have a similar likelihood in web search, whereas the former is by far the most likely aspect of the query *"bond"* when searching for images. As a result, a natural question when building aggregated search interfaces is how to tackle query ambiguity across multiple search verticals. To address this problem, Santos et al. [2011a] discussed possible extensions for state-of-the-art diversification approaches, as well as for the currently established evaluation paradigm, in order to accommodate per-vertical notions of query ambiguity, aspect likelihood, document coverage, and novelty.

6.4 Summary

In this chapter, we have surveyed advanced approaches for two broad problems closely related to search result diversification, namely, query ambiguity detection and query aspect mining. Besides reviewing the most prominent approaches for each of these problems, we discussed how to leverage the described solutions for an improved diversification effectiveness. In addition, we provided an overview of current applications of search result diversification beyond web search. In the next chapter, we will conclude this survey and provide a landscape of open research directions in the field of search result diversification.

7

Summary and Open Directions

While the ever-increasing rates of information production and consumption severely impact the design of efficient search engines, achieving an effective retrieval performance remains a challenge [Cutts, 2012]. In particular, despite the continued growth of the Web, typical web search queries are still short [Jansen et al., 2000]. As an immediate result, queries submitted to a web search engine are often ambiguous to some extent [Song et al., 2009]. As discussed throughout this survey, a central approach for tackling query ambiguity is to diversify the search results. In this chapter, we briefly summarise the main points covered in this survey, and provide a landscape of open research directions.

7.1 Summary of this Survey

In this survey, we described the major advances in the field of search result diversification over the past decades. In particular, in Chapter 1, we provided a historical background on the development of relevance-oriented ranking in information retrieval. In addition, we discussed the limitations of traditional ranking approaches in light of ambiguity and redundancy, as a motivation for diversifying these results. In Chapter 2, we provided a formal definition of the search result diversification problem and discussed its NP-hardness. Furthermore, we introduced a taxonomy for categorising existing diversification approaches in the literature according to two orthogonal dimensions, namely, aspect representation and diversification strategy. In Chapters 3 and 4, we introduced the most prominent diversification approaches in the literature that adopt implicit or explicit aspect representations, respectively. While the former have the longest history in the search result diversification literature, the latter represent the current state-of-the-art in the field. In Chapter 5, we provided a comprehensive account of the currently established paradigm for diversity evaluation. Besides describing existing benchmark test collections, we discussed several metrics for quantitatively evaluating diversity-oriented rankings. Lastly, in Chapter 6, we discussed recent advances in the field of search result diversification, primarily motivated by improvements in the related areas of query ambiguity detection and query aspect mining. In addition, we described several applications of diversification beyond web search.

7.2 Open Research Directions

Search result diversification has received a considerable attention from the IR community in recent years, with notable achievements both in the generation as well as in the evaluation of diversity-oriented rankings. Nevertheless, several directions for further research are still open. In this section, we discuss prominent open directions related to the modelling, estimation, and evaluation of diversification approaches.

7.2.1 Modelling

As discussed throughout Chapters 3 and 4, several ranking models have been proposed in the literature to produce diversity-oriented rankings, most of them directly built upon the greedy approximation approach described in Algorithm 2.1. Despite the approximation guarantees offered by this greedy formulation, an improved effectiveness could be attained by exploring fast exact solutions for the problem [Woeginger, 2003]. This is particularly motivated by the facts that ambiguous queries typically have very few dominant aspects and that diversification is commonly performed as a post-process, aimed at re-ranking a top few documents retrieved by a relevance-oriented approach.

Besides the underlying optimisation algorithm, the optimisation objective plays a central role in the effectiveness of any diversification approach. In turn, an improved modelling of such an optimisation objective could be pursued in several directions. For instance, a common assumption made by existing diversification approaches is that the identified query aspects are independent and identically distributed (IID) with respect to one another. On the one hand, such an assumption greatly simplifies the underlying mathematics of these approaches and their practical instantiation. On the other hand, this assumption is unlikely to hold in a real scenario, particularly for aspects underlying an underspecified query [Clarke et al., 2008]. Indeed, the rankings produced for such aspects tend to be highly correlated, as the representation of these aspects (e.g., as query suggestions) often share common terms [Ma et al., 2010]. From a machine learning perspective, one possibility is to account for partial dependencies between candidate aspects associated with the same query [Dundar et al., 2007].

In addition to the interplay between different aspects underlying the same query, an open direction for investigation is an effective modelling of the interplay between the retrieved documents. In particular, despite numerous attempts to model dependences between the retrieved documents, as introduced in §3.1, it remains to be shown whether novelty per se could be an effective diversification strategy. Indeed, pure novelty-based approaches have been shown to underperform in a web search setting, whereas hybrid approaches promoting both coverage and novelty have only exhibited a marginal edge over pure coverage-based ones [Santos et al., 2012b]. Further analyses have shown that, compared to coverage, novelty is a much riskier diversification strategy, hence exhibiting an inconsistent effectiveness across different queries [Santos, 2013]. As a result, a promising direction for improving existing novelty-based and hybrid approaches is to promote novelty selectively, by modelling the amount of redundancy (and hence the need for novelty) permeating the results of each individual query.

Besides the interplay between aspects and between documents, diversification involves modelling the interplay across aspects and documents. While most current approaches model each of these interplays separately, it would be desirable to have a unified process for learning a diversification model given suitable training data. Indeed, learning to rank for diversity is a recognised open challenge also in the machine learning community [Chapelle et al., 2011a]. As described in Chapter 3, existing approaches for learning to diversify are based on implicit aspect representations [Yue and Joachims, 2008, Radlinski et al., 2008, Slivkins et al., 2010, Raman et al., 2012, Zhu et al., 2014]. A machine-learned explicit diversification approach that fully explores the connections between and across query aspects and documents is missing. Such an approach could unify the processes of weighing both the relative importance of different aspects as well as the retrieved documents' coverage and novelty with respect to each aspect. Moreover, it could help integrate further features related to the user's preferences to enable a personalised diversification [Vallet and Castells, 2012], as well as contextual features, such as location and time [Nguyen and Kanhabua, 2014], to enable a context-aware diversification [Melucci, 2012].

7.2.2 Estimation

The search result diversification problem can be decomposed into three interrelated problems, namely, query ambiguity detection, query aspect mining, and document ranking. As discussed in Chapter 6, several advanced solutions have been recently introduced for each of these problems. Regarding query ambiguity detection, initial results have demonstrated the potential of selectively adapting the amount of diversification for the predicted ambiguity of each individual query [Santos et al., 2010a]. More advanced query classification approaches could be considered to improve the accuracy of ambiguity detection. Furthermore, the resulting selective approaches could be further deployed to improve the robustness of diversification across different queries. Indeed, as discussed in §7.2.1, current novelty-based approaches are particularly risky. Nonetheless, even pure coverage-based and hybrid approaches could benefit from an improved robustness.

Regarding query aspect mining, query reformulations mined from a query log generally provide the most effective aspect representation, more closely modelling real users' information needs [Kruschwitz et al., 2013]. On the other hand, a substantial fraction of queries have very sparse or no past usage in a query log [Downey et al., 2007], which may hamper the effectiveness of search result diversification approaches based on query reformulations. Tackling such cold-start queries is a challenging problem. Besides attempts to identify query aspects from alternative sources, such as anchor-text [Dou et al., 2011], promising directions include a deeper understanding of the query, by leveraging taxonomies [Agrawal et al., 2009], thesauri [Plakhov, 2011], or knowledge bases [Zheng et al., 2012], and by analysing structural components of the query [Dang and Croft, 2013]. In the same vein, generating effective implicit aspect representations is also an open problem. A promising direction towards a pure implicit aspect generation is a supervised approach aimed at learning the characteristics of effective aspects given only the top retrieved documents. As discussed in §6.2, aspect mining approaches inspired by frequent pattern extraction [Zheng et al., 2011b] and query expansion [Bouchoucha et al., 2013, Vargas et al., 2013] have shown promise.

From the perspective of document ranking, two central questions must be addressed: how to weigh the relative importance of different aspects, and how to estimate the relevance of each document with respect to each aspect. Regarding the former, a few works have tried to estimate the importance of a given aspect according to how much it is covered in the target collection [Santos et al., 2010b,c]. Nevertheless, assuming that all aspects are equally uniform has usually yielded more effective results. An alternative to such content-based popularity approaches is to directly estimate the importance of each aspect as perceived by the population of users (as opposed to the population of content producers). This is in line with recent approaches that seek to learn to rank query reformulations from a query log to attain an improved diversification performance [Santos et al., 2013]. Another important consideration is whether the inferred importance of different aspects is respected in the diversified ranking. In particular, Ozdemiray and Altingovde [2013] noted that existing approaches may inadvertently consider an aspect as being satisfied too early. One possibility for addressing such an "aspect fading" is to enforce a proportional coverage of the identified aspects [Dang and Croft, 2012].

The question of how to estimate the relevance of each document with respect to each aspect is central to the estimation of coverage and novelty. Several approaches have been proposed to this end, from using classification confidence scores [Agrawal et al., 2009] or topic modelling probabilities [Carterette and Chandar, 2009] as proxies for relevance, to using traditional relevance-oriented ranking models [Santos et al., 2010b] or advanced machine-learned ranking models to infer relevance with respect to the intent underlying each aspect [Santos et al., 2011b]. In line with the discussion in §7.2.1 regarding a unified model for diversification, a possible direction for further improving the document ranking component of existing approaches is a mechanism to combine the estimation of aspect importance and aspect-level document relevance in an iterative learning process. Such a process could optimise the selection of query aspects and of retrieved documents in subsequent iterations, with the output of either optimisation serving as input to the other, until converging into stable rankings. As a result, aspect mining and diversification could be tackled simultaneously.

7.2.3 Evaluation

The development of evaluation metrics that better reflect the behaviour of real search users is a major concern for retrieval evaluation in general, as well as for search result diversification in particular. On the other hand, two additional challenges are particularly relevant when evaluating diversity-oriented rankings, namely, the reusability of existing test collections and the cost of building new ones. As discussed in §5.3, the reusability of the currently available test collections for diversity evaluation is compromised by their relatively shallow assessment pools compared with the web-scale document corpora from where these pools were originally sampled. While condensed-list evaluation metrics have been suggested as a means to attenuate the problem [Sakai, 2007], in order to achieve full reusability, additional relevance assessments are needed whenever there are unjudged documents [Carterette, 2007]. A possible direction towards this end is to develop suitable evaluation protocols for performing large-scale user-centred relevance assessments through crowdsourcing [Vallet, 2011].

Building benchmark test collections for diversity evaluation and keeping them reusable by performing additional assessments can be costly. This is particularly due to the requirements of identifying ground-truth query aspects a priori and of performing relevance assessments at the aspect level. Hence, relaxing these requirements while ensuring a meaningful evaluation is a promising direction for investigation. Regarding the requirement of identifying ground-truth aspects a priori, Sakai [2013] called for approaches that could automatically identify meaningful aspects given a target set of rankings to be assessed. While forgoing an assessment of absolute diversity, such approaches would still allow for assessing the relative diversity of different rankings. However, Sakai et al. [2013a] noted that even the relative comparison of rankings may be affected by a particular choice of aspects. Regarding the requirement of assessing relevance at the aspect level, a completely revamped evaluation paradigm would be needed. A possible direction here is to rely on multiple assessors to judge the relevance of the documents retrieved for any given query. In this vein, as discussed in §5.2, the approach introduced by Chandar and Carterette [2013] based on pairwise preferences shows promise.

7.3 Concluding Remarks

In this chapter, we have briefly summarised the materials discussed throughout this survey. In addition, we have provided a landscape of open research directions concerning the modelling, estimation, and evaluation of diversification approaches. We believe that the quantity and quality of the works described in this survey testify to the maturity of search result diversification as a research field. At the same time, the steady pace of publications in the field appearing at top information retrieval venues together with the open research directions identified in this chapter are an indication of a promising future ahead.

References

R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, pages 5–14, Barcelona, Spain, 2009. ACM.

G. Amati. Frequentist and Bayesian approach to information retrieval. In *Proceedings of the 28th European Conference on IR Research on Advances in Information Retrieval*, pages 13–24, London, UK, 2006. Springer.

G. Amati. *Probability models for information retrieval based on Divergence From Randomness*. PhD thesis, University of Glasgow, 2003.

A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the Web. *ACM Transactions on Internet Technology*, 1(1):2–43, 2001. ISSN 1533-5399.

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

J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34, Salvador, Brazil, 2005. ACM.

R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query recommendation using query logs in search engines. In *Proceedings of the 9th International Conference on Current Trends in Database Technology*, pages 588–596, Heraklion, Greece, 2004. Springer-Verlag.

R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Pearson Education Ltd., Harlow, UK, 2 edition, 2011.

D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 407–416, Boston, MA, USA, 2000. ACM.

F. Belém, R. L. T. Santos, J. Almeida, and M. A. Gonçalves. Topic diversity in tag recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 141–148, Hong Kong, China, 2013. ACM.

Y. Bernstein and J. Zobel. Redundant documents and search effectiveness. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 736–743, Bremen, Germany, 2005. ACM.

D. Berry and B. Fristedt. *Bandit problems: Sequential allocation of experiments*. Chapman and Hall, 1985.

S. Bhatia, D. Majumdar, and P. Mitra. Query suggestions in the absence of query logs. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 795–804, Beijing, China, 2011. ACM.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003. ISSN 1532-4435.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 2008 (10):P10008+, 2008. ISSN 1742-5468.

P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 609–618, Napa Valley, CA, USA, 2008. ACM.

P. Boldi, F. Bonchi, C. Castillo, D. Donato, and S. Vigna. Query suggestions using query-flow graphs. In *Proceedings of the 2009 Workshop on Web Search Click Data*, pages 56–63. ACM, 2009a.

P. Boldi, F. Bonchi, C. Castillo, and S. Vigna. From "Dango" to "Japanese cakes": Query reformulation models and patterns. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 183–190, Milan, Italy, 2009b. IEEE Computer Society.

P. Borlund. The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10):913–925, 2003. ISSN 1532-2882.

A. Bouchoucha, J. He, and J.-Y. Nie. Diversified query expansion using ConceptNet. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pages 1861–1864, San Francisco, CA, USA, 2013. ACM.

D. Broccolo, L. Marcon, F. M. Nardini, R. Perego, and F. Silvestri. Generating suggestions for queries in the long tail with an inverted index. *Information Processing and Management*, 48(2):326–339, 2012. ISSN 0306-4573.

A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002. ISSN 0163-5840.

C. Buckley. Why current IR engines fail. In *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval*, pages 584–585, Sheffield, UK, 2004. ACM Press.

G. Capannini, F. M. Nardini, R. Perego, and F. Silvestri. Efficient diversification of web search results. *Proceedings of the VLDB Endowment*, 4(7): 451–459, 2011. ISSN 2150-8097.

J. Carbonell and J. 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 Informaion Retrieval*, pages 335–336, Melbourne, Australia, 1998. ACM.

B. Carterette. Robust test collections for retrieval evaluation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55–62, Amsterdam, The Netherlands, 2007. ACM.

B. Carterette. An analysis of NP-completeness in novelty and diversity ranking. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval*, pages 200–211, Cambridge, UK, 2009. Springer-Verlag.

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

P. Chandar and B. 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*, pages 413–422, Dublin, Ireland, 2013. ACM.

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

O. Chapelle, Y. Chang, and T.-Y. Liu. Future directions in learning to rank. *Journal of Machine Learning Research, Proceedings Track*, pages 91–100, 2011a.

O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. Intent-based diversification of web search results: Metrics and algorithms. *Information Retrieval*, 14(6):572–592, 2011b.

H. 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, Seattle, WA, USA, 2006. ACM.

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

C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web track. In *Proceedings of the 18th Text REtrieval Conference*, Gaithersburg, MD, USA, 2009a.

C. L. A. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval*, pages 188–199, Cambridge, UK, 2009b. Springer-Verlag.

C. L. A. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the TREC 2010 Web track. In *Proceedings of the 19th Text REtrieval Conference*, Gaithersburg, MD, USA, 2010.

C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 75–84, Hong Kong, China, 2011a. ACM.

C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. Overview of the TREC 2011 Web track. In *Proceedings of the 20th Text REtrieval Conference*, Gaithersburg, MD, USA, 2011b.

C. L. A. Clarke, N. Craswell, and E. M. Voorhees. Overview of the TREC 2012 Web track. In *Proceedings of the 21st Text REtrieval Conference*, Gaithersburg, MD, USA, 2012.

C. Cleverdon. The Cranfield tests on index language devices. *Aslib Proceedings*, 19(6):173–194, 1967.

C. W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, Chicago, IL, USA, 1991. ACM.

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

K. Collins-Thompson, P. Bennett, F. Diaz, C. L. A. Clarke, and E. M. Voorhees. TREC 2013 Web track overview. In *Proceedings of the 22nd Text REtrieval Conference*, Gaithersburg, MD, USA, 2013.

W. S. Cooper. The inadequacy of probability of usefulness as a ranking criterion for retrieval system output. Technical report, University of California, Berkeley, Berkeley, CA, USA, 1971.

W. S. Cooper. Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. *ACM Transactions on Information Systems*, 13(1):100–111, 1995.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2nd edition, 2001.

N. Craswell and M. Szummer. Random walks on the click graph. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 239–246, Amsterdam, The Netherlands, 2007. ACM.

N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 1st International Conference on Web Search and Data Mining*, pages 87–94. ACM, 2008.

S. Cronen-Townsend and W. B. Croft. Quantifying query ambiguity. In *Proceedings of the 2nd International Conference on Human Language Technology Research*, pages 104–109, San Diego, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

M. Cutts. Spotlight keynote. In *Proceedings of Search Engine Strategies*, San Francisco, CA, USA, 2012.

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

V. Dang and W. B. Croft. Query reformulation using anchor text. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pages 41–50, New York, NY, USA, 2010. ACM.

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

V. Dang, X. Xue, and W. B. Croft. Inferring query aspects from reformulations using clustering. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 2117–2120, Glasgow, UK, 2011. ACM.

G. Demartini. ARES: a retrieval engine based on sentiments sentiment-based search result annotation and diversification. In *Proceedings of the 33rd European Conference on IR Research on Advances in Information Retrieval*, pages 772–775, Dublin, Ireland, 2011. Springer-Verlag.

E. Demidova, P. Fankhauser, X. Zhou, and W. Nejdl. Divq: Diversification for keyword search over structured databases. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 331–338, Geneva, Switzerland, 2010. ACM.

T. Deselaers, T. Gass, P. Dreuw, and H. Ney. Jointly optimising relevance and diversity in image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 1–8, Santorini, Greece, 2009. ACM.

F. Diaz, M. Lalmas, and M. Shokouhi. From federated to aggregated search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 910, 2010.

Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proceedings of the fourth ACM international Conference on Web Search and Data Mining*, pages 475–484, Hong Kong, China, 2011. ACM.

D. Downey, S. Dumais, and E. Horvitz. Heads and tails: studies of web search with common and rare queries. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 847–848, Amsterdam, The Netherlands, 2007. ACM.

M. Dundar, B. Krishnapuram, J. Bi, and R. B. Rao. Learning classifiers when the training data is not IID. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 756–761, Hyderabad, India, 2007. Morgan Kaufmann Publishers Inc.

N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 459–460, Toronto, Canada, 2003. ACM.

U. Feige. A threshold of ln(n) for approximating set cover. *Journal of the ACM*, 45:634–652, 1998. ISSN 0004-5411.

B. M. Fonseca, P. B. Golgher, E. S. De Moura, B. Pôssas, and N. Ziviani. Discovering search engine related queries using association rules. *Journal of Web Engineering*, 2(4):215–227, October 2003. ISSN 1540-9589.

E. A. Fox and J. A. Shaw. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference*, pages 243–252, Gaithersburg, MD, USA, 1993.

J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

X. Geng, T.-Y. Liu, T. Qin, A. Arnold, H. Li, and H.-Y. Shum. Query dependent ranking using k-nearest neighbor. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–122, Singapore, Singapore, 2008. ACM.

V. Gil-Costa, R. L. T. Santos, C. Macdonald, and I. Ounis. Sparse spatial selection for novelty-based search result diversification. In *Proceedings of the 18th International Symposium on String Processing and Information Retrieval*, pages 344–355, Pisa, Italy, 2011. Springer.

V. Gil-Costa, R. L. T. Santos, C. Macdonald, and I. Ounis. Modelling efficient novelty-based search result diversification in metric spaces. *Journal of Discrete Algorithms*, 18:75–88, 2013. ISSN 1570-8667.

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

P. B. Golbus, J. A. Aslam, and C. L. Clarke. Increasing evaluation sensitivity to diversity. *Information Retrieval*, 16(4):530–555, 2013. ISSN 1386-4564.

S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *Proceedings of the 18th International Conference on World Wide Web*, pages 381–390, Madrid, Spain, 2009. ACM.

M. D. Gordon and P. Lenk. A utility theoretic examination of the probability ranking principle in information retrieval. *Journal of the American Society for Information Science and Technology*, 42(10):703–714, 1991.

M. D. Gordon and P. Lenk. When is the probability ranking principle suboptimal? *Journal of the American Society for Information Science and Technology*, 43(1):1–14, 1992.

D. Harman. Overview of the second Text REtrieval Conference (TREC-2). In *Proceedings of the 2nd Text REtrieval Conference*, Gaithersburg, MD, USA, 1993.

S. P. Harter. A probabilistic approach to automatic keyword indexing. Part I: On the distribution of specialty words in a technical literature. *Journal of the American Society for Information Science*, 26(4):197–206, 1975a.

S. P. Harter. A probabilistic approach to automatic keyword indexing. Part II: An algorithm for probabilistc indexing. *Journal of the American Society for Informaiton Science*, 26(4):280–289, 1975b.

J. He, E. Meij, and M. de Rijke. Result diversification based on query-specific cluster ranking. *Journal of the American Society for Information Science and Technology*, 62(3):550–571, 2011. ISSN 1532-2882.

J. He, V. Hollink, and A. 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*, pages 851–860, Portland, OR, USA, 2012. ACM.

W. R. Hersh and P. Over. TREC-8 Interactive track report. In *Proceedings of the 8th Text REtrieval Conference*, Gaithersburg, MD, USA, 1999.

D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*, pages 569–584, Heraklion, Greece, 1998. Springer.

D. S. Hochbaum, editor. *Approximation algorithms for NP-hard problems*. PWS Publishing Co., Boston, MA, USA, 1997.

D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender Systems: An Introduction*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.

B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the Web. *SIGIR Forum*, 32(1):5–17, 1998. ISSN 0163-5840.

B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing and Management*, 36(2):207–227, 2000. ISSN 0306-4573.

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. ISSN 1046-8188.

R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *Proceedings of the 15th international conference on World Wide Web*, pages 387–396, Edinburgh, UK, 2006. ACM.

I.-H. Kang and G. Kim. Query type classification for web document retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 64–71, Toronto, Canada, 2003. ACM.

J. G. Kemeny and J. L. Snell. *Finite Markov Chains.* Springer, 1960.

S. Kharazmi, M. Sanderson, F. Scholer, and D. Vallet. Using score differences for search result diversification. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1143–1146. ACM, 2014.

E. Kharitonov, C. Macdonald, P. Serdyukov, and I. Ounis. Intent models for contextualising and diversifying query suggestions. In *Proceedings of the 22nd ACM International Conference on Conference on information and Knowledge Management*, pages 2303–2308, San Francisco, CA, USA, 2013. ACM.

S. Khuller, A. Moss, and J. S. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70:39–45, 1999. ISSN 0020-0190.

Y. Kim and W. B. Croft. Diversifying query suggestions based on query documents. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 891–894, Gold Coast, QLD, Australia, 2014. ACM.

R. Kraft and J. Zien. Mining anchor text for query refinement. In *Proceedings of the 13th International Conference on World Wide Web*, pages 666–674, New York, NY, USA, 2004. ACM.

R. Krestel and N. Dokoohaki. Diversifying product review rankings: Getting the full picture. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pages 138–145, Washington, DC, USA, 2011. IEEE Computer Society.

U. Kruschwitz, D. Lungley, M.-D. Albakour, and D. Song. Deriving query suggestions for site search. *Journal of the American Society for Information Science and Technology*, 64(10):1975–1994, 2013. ISSN 1532-2890.

E. Lagergren and P. Over. Comparing interactive information retrieval systems across sites: The TREC-6 Interactive track matrix experiment. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 164–172, Melbourne, Australia, 1998. ACM.

N. Lathia, S. Hailes, L. Capra, and X. Amatriain. Temporal diversity in recommender systems. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 210–217, Geneva, Switzerland, 2010. ACM.

V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 120–127, New Orleans, LA, USA, 2001. ACM.

T. Leelanupab, G. Zuccon, and J. M. Jose. A comprehensive analysis of parameter settings for novelty-biased cumulative gain. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1950–1954, Maui, HI, USA, 2012. ACM.

S. Liang, Z. Ren, and M. de Rijke. Fusion helps diversification. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 303–312, Gold Coast, QLD, Australia, 2014. ACM.

N. Limsopatham, C. Macdonald, and I. Ounis. Modelling relevance towards multiple inclusion criteria when ranking patients. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, pages 1639–1648, Shanghai, China, 2014. ACM.

H. Liu and P. Singh. ConceptNet—a practical commonsense reasoning toolkit. *BT Technology Journal*, 22(4):211–226, 2004. ISSN 1358-3948.

H. Ma, M. R. Lyu, and I. King. Diversifying query suggestion results. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, Atlanta, GA, USA, 2010. AAAI Press.

J. I. Marden. *Analyzing and modeling rank data*. Taylor & Francis, 1996.

H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. ISSN 00221082.

M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7(3):216–244, 1960. ISSN 0004-5411.

Q. Mei, D. Zhou, and K. Church. Query suggestion using hitting time. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 469–478, Napa Valley, CA, USA, 2008. ACM.

M. Melucci. Contextual search: A computational framework. *Foundations and Trends in Information Retrieval*, 6(4-5):257–405, 2012.

S. Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997. ISSN 0002-8231.

A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):1–27, 2008. ISSN 1046-8188.

V. Murdock and M. Lalmas. Workshop on aggregated search. *SIGIR Forum*, 42:80–83, 2008. ISSN 0163-5840.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14:265–294, 1978. ISSN 0025-5610.

T. N. Nguyen and N. Kanhabua. Leveraging dynamic query subtopics for time-aware search result diversification. In *Proceedings of the 36th European Conference on Information Retrieval*, pages 222–234, Amsterdam, The Netherlands, 2014. Springer.

P. Over. TREC-6 Interactive report. In *Proceedings of the 6th Text REtrieval Conference*, pages 73–81, Gaithersburg, MD, USA, 1997.

P. Over. TREC-7 Interactive track report. In *Proceedings of the 7th Text REtrieval Conference*, pages 33–39, Gaithersburg, MD, USA, 1998.

A. M. Ozdemiray and I. S. Altingovde. Score and rank aggregation methods for explicit search result diversification. Technical Report METU-CENG-2013-01, Middle East Technical University, Ankara, Turkey, 2013.

M. L. Paramita, J. Tang, and M. Sanderson. Generic and spatial approaches to image search results diversification. In *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval*, pages 603–610, Toulouse, France, 2009. Springer.

J. Peng, C. Macdonald, and I. Ounis. Learning to select a ranking function. In *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval*, pages 114–126, Milton Keynes, UK, 2010. Springer.

V. Plachouras. Diversity in expert search. In *Proceedings of the 1st International Workshop on Diversity in Document Retrieval*, pages 63–67, Dublin, Ireland, 2011.

A. Plakhov. Entity-oriented search result diversification. In *Proceedings of the 1st International Workshop on Entity-Oriented Search*, Beijing, China, 2011.

J. M. Ponte and W. B. 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*, pages 275–281, Melbourne, Australia, 1998. ACM.

F. Radlinski and S. 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*, pages 691–692, Seattle, WA, USA, 2006. ACM.

F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning*, pages 784–791, Helsinki, Finland, 2008. ACM.

F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims. Redundancy, diversity and interdependent document relevance. *SIGIR Forum*, 43(2): 46–52, 2009. ISSN 0163-5840.

F. Radlinski, M. Szummer, and N. Craswell. Metrics for assessing sets of subtopics. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 853–854, Geneva, Switzerland, 2010a. ACM.

F. Radlinski, M. Szummer, and N. Craswell. Inferring query intent from reformulations and clicks. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1171–1172, Raleigh, NC, USA, 2010b.

D. Rafiei, K. Bharat, and A. Shukla. Diversifying web search results. In *Proceedings of the 19th International Conference on World Wide Web*, pages 781–790, Raleigh, NC, USA, 2010.

K. Raman, P. Shivaswamy, and T. Joachims. Online learning to diversify from implicit feedback. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, China, 2012. ACM.

S. E. Robertson and K. Spärck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1976.

S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter. Probabilistic models of indexing and searching. In *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*, pages 35–56. Butterworth & Co., 1981.

S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. ISSN 1554-0669.

S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, pages 42–49, Washington, DC, USA, 2004. ACM.

S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*, Gaithersburg, MD, USA, 1994.

D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web*, pages 13–19, New York, NY, USA, 2004. ACM.

B. R. Rowe, D. W. Wood, A. N. Link, and D. A. Simoni. Economic impact assessment of NIST's Text REtrieval Conference (TREC) program. Technical Report 0211875, RTI International, 2010.

T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 525–532, Seattle, WA, USA, 2006. ACM.

T. Sakai. Alternatives to bpref. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 71–78, Amsterdam, The Netherlands, 2007. ACM.

T. Sakai. Evaluation with informational and navigational intents. In *Proceedings of the 21st International Conference on World Wide Web*, pages 499–508, Lyon, France, 2012. ACM.

T. Sakai. The unreusability of diversified search test collections. In *Proceedings of the 5th International Workshop on Evaluating Information Access*, pages 1–8, Tokyo, Japan, 2013.

T. Sakai and R. Song. Diversified search evaluation: Lessons from the NTCIR-9 Intent task. *Information Retrieval*, 2012. ISSN 1386-4564.

T. Sakai, N. Craswell, R. Song, S. 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*, pages 42–50, Tokyo, Japan, 2010. NII.

T. Sakai, Z. Dou, and C. L. Clarke. The impact of intent selection on diversified search evaluation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 921–924, Dublin, Ireland, 2013a. ACM.

T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, and R. Song. Overview of the NTCIR-10 Intent-2 task. In *Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, Tokyo, Japan, 2013b.

P. A. Samuelson and W. D. Nordhaus. *Microeconomics*. McGraw-Hill, 2001.

M. 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*, pages 499–506, Singapore, Singapore, 2008. ACM.

M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

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

R. L. T. Santos. *Explicit web search result diversification*. PhD thesis, School of Computing Science, University of Glasgow, Glasgow, UK, 2013.

R. L. T. Santos and I. Ounis. Diversifying for multiple information needs. In *Proceedings of the 1st International Workshop on Diversity in Document Retrieval*, pages 37–41, Dublin, Ireland, 2011.

R. L. T. Santos, C. Macdonald, and I. Ounis. Selectively diversifying web search results. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1179–1188, Toronto, Canada, 2010a. ACM.

R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web*, pages 881–890, Raleigh, NC, USA, 2010b. ACM.

R. L. T. Santos, J. Peng, C. Macdonald, and I. Ounis. Explicit search result diversification through sub-queries. In *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval*, pages 87–99, Milton Keynes, UK, 2010c. Springer.

R. L. T. Santos, C. Macdonald, and I. Ounis. Aggregated search result diversification. In *Proceedings of the 3rd International Conference on the Theory of Information Retrieval*, pages 250–261, Bertinoro, Italy, 2011a. Springer.

R. L. T. Santos, C. Macdonald, and I. Ounis. Intent-aware search result diversification. In *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 595–604, Beijing, China, 2011b. ACM.

R. L. T. Santos, C. Macdonald, R. McCreadie, I. Ounis, and I. Soboroff. Information retrieval on the blogosphere. *Foundations and Trends in Information Retrieval*, 6(1):1–125, 2012a.

R. L. T. Santos, C. Macdonald, and I. Ounis. On the role of novelty for search result diversification. *Information Retrieval*, 15(5):478–502, 2012b.

R. L. T. Santos, C. Macdonald, and I. Ounis. Learning to rank query suggestions for adhoc and diversity search. *Information Retrieval*, 16(4):429–451, 2013. ISSN 1386-4564.

M. Searcóid. *Metric Spaces.* Springer Undergraduate Mathematics Series. Springer, 2006.

C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999. ISSN 0163-5840.

F. Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4(1-2):1–174, 2010.

A. Slivkins, F. Radlinski, and S. Gollapudi. Learning optimally diverse rankings over large document collections. In *Proceedings of the 27th Annual International Conference on Machine Learning*, pages 983–990, Haifa, Israel, 2010. Omnipress.

R. Song, Z. Luo, J.-Y. Nie, Y. Yu, and H.-W. Hon. Identification of ambiguous queries in web search. *Information Processing and Management*, 45(2):216–229, 2009. ISSN 0306-4573.

R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the NTCIR-9 Intent task. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, Tokyo, Japan, 2011a.

Y. Song, D. Zhou, and L. wei He. Post-ranking query suggestion by diversifying search results. In *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 815–824, Beijing, China, 2011b. ACM.

K. Spärck-Jones, S. E. Robertson, and M. Sanderson. Ambiguous requests: Implications for retrieval tests, systems and theories. *SIGIR Forum*, 41(2): 8–17, 2007. ISSN 0163-5840.

I. Szpektor, A. Gionis, and Y. Maarek. Improving recommendation for long-tail queries via templates. In *Proceedings of the 20th international conference on World wide web*, pages 47–56, Hyderabad, India, 2011. ACM.

J. Teevan, S. T. Dumais, and E. Horvitz. Characterizing the value of personalizing search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 757–758, Amsterdam, The Netherlands, 2007. ACM.

I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005. ISSN 1532-4435.

H. R. Turtle and W. B. Croft. Uncertainty in information retrieval systems. In *Uncertainty Management in Information Systems*, pages 189–224. Kluwer Academic Publishers, Norwell, MA, USA, 1996.

D. Vallet. Crowdsourced evaluation of personalization and diversification techniques in web search. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, Beijing, China, 2011. ACM.

D. Vallet and P. Castells. Personalized diversification of search results. In *Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 841–850, Portland, OR, USA, 2012. ACM.

R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *Proceedings of the 18th International Conference on World Wide Web*, pages 341–350, Madrid, Spain, 2009. ACM.

C. J. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, New York, NY, USA, 2004.

S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems*, pages 109–116, Chicago, IL, USA, 2011. ACM.

S. Vargas, P. Castells, and D. Vallet. Intent-oriented diversity in recommender systems. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1211–1212, Beijing, China, 2011. ACM.

S. Vargas, R. L. T. Santos, C. Macdonald, and I. Ounis. Selecting effective expansion terms for diversity. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 69–76, Lisbon, Portugal, 2013. CID.

E. Vee, U. Srivastava, J. Shanmugasundaram, P. Bhat, and S. A. Yahia. Efficient computation of diverse query results. In *Proceedings of the 24th International Conference on Data Engineering*, pages 228–236, Cancún, Mexico, 2008. IEEE Computer Society.

R. V. Vohra and N. G. Hall. A probabilistic analysis of the maximal covering location problem. *Discrete Applied Mathematics*, 43(2):175–183, 1993. ISSN 0166-218X.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, 1944.

E. M. Voorhees. TREC: Continuing information retrieval's tradition of experimentation. *Communications of the ACM*, 50(11):51–54, 2007. ISSN 0001-0782.

E. M. Voorhees and D. Harman. Overview of the 6th Text REtrieval Conference. In *Proceedings of the 6th Text REtrieval Conference*, Gaithersburg, MD, USA, 1997.

E. M. Voorhees and D. Harman. Overview of the 7th Text REtrieval Conference. In *Proceedings of the 7th Text REtrieval Conference*, Gaithersburg, MD, USA, 1998.

E. M. Voorhees and D. Harman. Overview of the 8th Text REtrieval Conference. In *Proceedings of the 8th Text REtrieval Conference*, Gaithersburg, MD, USA, 1999.

E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval.* Digital Libraries and Electronic Publishing. MIT Press, 2005.

J. Wang and J. 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, Boston, MA, USA, 2009. ACM.

Q. Wang, Y. Qian, R. Song, Z. Dou, F. Zhang, T. Sakai, and Q. Zheng. Mining subtopics from text fragments for a web query. *Information Retrieval*, 16 (4):484–503, 2013. ISSN 1386-4564.

X. Wang and C. Zhai. Mining term association patterns from search logs for effective query reformulation. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 479–488, Napa Valley, CA, USA, 2008. ACM.

X. Wang, H. Fang, and C. Zhai. A study of methods for negative relevance feedback. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 219–226, Singapore, Singapore, 2008. ACM.

M. J. Welch, J. Cho, and C. Olston. Search result diversity for informational queries. In *Proceedings of the 20th International Conference on World Wide Web*, pages 237–246, Hyderabad, India, 2011. ACM.

G. J. Woeginger. Exact algorithms for NP-hard problems: A survey. In *Combinatorial Optimization—Eureka, You Shrink!*, pages 185–207. Springer, 2003.

M. A. Woodbury. Inverting modified matrices. Technical Report MR38136, Statistical Research Group, Princeton University, Princeton, NJ, USA, 1950.

F. Wu, J. Madhavan, and A. Halevy. Identifying aspects for web-search queries. *Journal of Artificial Intelligence Research*, 40(1):677–700, 2011. ISSN 1076-9757.

X. Yin, J. X. Huang, X. Zhou, and Z. Li. A survival modeling approach to biomedical search result diversification using Wikipedia. In *Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 901–902, Geneva, Switzerland, 2010. ACM.

C. Yu, L. Lakshmanan, and S. Amer-Yahia. It takes variety to make a world: Diversification in recommender systems. In *Proceedings of the 12th International Conference on Extending Database Technology*, pages 368–378, Saint Petersburg, Russia, 2009. ACM.

Y. Yue and T. Joachims. Predicting diverse subsets using structural svms. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1224–1231, Helsinki, Finland, 2008. ACM.

H. Zaragoza, N. Craswell, M. J. Taylor, S. Saria, and S. E. Robertson. Microsoft Cambridge at TREC 13: Web and Hard tracks. In *Proceedings of the 13th Text REtrieval Conference*, Gaithersburg, MD, USA, 2004.

C. Zhai. Statistical language models for information retrieval: A critical review. *Foundations and Trends in Information Retrieval*, 2(3):137–213, 2008. ISSN 1554-0669.

C. Zhai and J. Lafferty. A risk minimization framework for information retrieval. *Information Processing and Management*, 42(1):31–55, 2006. ISSN 0306-4573.

C. Zhai, W. W. Cohen, and J. 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, Toronto, Canada, 2003. ACM.

Z. Zhang and O. Nasraoui. Mining search engine query logs for query recommendation. In *Proceedings of the 15th international conference on World Wide Web*, pages 1039–1040, Edinburgh, UK, 2006. ACM.

W. Zheng, H. Fang, C. Yao, and M. Wang. Search result diversification for enterprise data. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1901–1904, Glasgow, UK, 2011a. ACM.

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

W. Zheng, H. Fang, and C. Yao. Exploiting concept hierarchy for result diversification. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1844–1848, Maui, HI, USA, 2012. ACM.

W. Zheng, H. Fang, C. Yao, and M. Wang. Leveraging integrated information to extract query subtopics for search result diversification. *Information Retrieval*, 17(1):52–73, 2014. ISSN 1386-4564.

T. Zhou, Z. Kuscsik, J. Liu, M. Medo, J. Wakeling, and Y. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.

X. Zhu, A. B. Goldberg, J. V. Gael, and D. Andrzejewski. Improving diversity in ranking using absorbing randomwalks. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics—Human Language Technologies*, pages 97–104, Rochester, NY, USA, 2007. ACL.

Y. Zhu, Y. Lan, J. Guo, X. Cheng, and S. Niu. Learning for search result diversification. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 293–302, Gold Coast, QLD, Australia, 2014. ACM.

C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, pages 22–32, Chiba, Japan, 2005. ACM.

J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, 1998. ACM.

G. Zuccon and L. Azzopardi. Using the quantum probability ranking principle to rank interdependent documents. In *Proceedings of the 32nd European Conference on IR Research on Advances in Information Retrieval*, pages 357–369, Milton Keynes, UK, 2010. Springer.

G. Zuccon, L. Azzopardi, D. Zhang, and J. Wang. Top-k retrieval using facility location analysis. In *Proceedings of the 34th European Conference on Advances in Information Retrieval*, pages 305–316, Barcelona, Spain, 2012. Springer-Verlag.