Foundations and Trends® in
Information Retrieval
9:1

Search Result Diversification

Rodrygo L. T. Santos, Craig Macdonald,
and Iadh Ounis

now
the essence of knowledge

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

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

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.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-000-2

Search Result Diversification

Foundations and Trends® in Information Retrieval
Volume 9, Issue 1, 2015

Editorial Scope

Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation, and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2015, Volume 9, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

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

DOI: 10.1561/1500000040

now
the essence of knowledge

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

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

Contents

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`