# Aggregated Search

**Jaime Arguello**
School of Information and Library Science
University of North Carolina at Chapel Hill, United States
jarguello@unc.edu

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

J. Arguello. *Aggregated Search*. Foundations and Trends® in Information Retrieval, vol. 10, no. 5, pp. 365–502, 2016.

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

ISBN: 978-1-68083-252-5

**Foundations and Trends® in Information Retrieval**
Volume 10, Issue 5, 2016

# 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, 2016, Volume 10, 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. 10, No. 5 (2016) 365–502

DOI: 10.1561/1500000052

# Aggregated Search

Jaime Arguello
School of Information and Library Science
University of North Carolina at Chapel Hill, United States
jarguello@unc.edu

# Contents

## Abstract

The goal of aggregated search is to provide integrated search across multiple heterogeneous search services in a unified interface—a single query box and a common presentation of results. In the web search domain, aggregated search systems are responsible for integrating results from specialized search services, or verticals, alongside the core web results. For example, search portals such as Google, Bing, and Yahoo! provide access to vertical search engines that focus on different types of media (images and video), different types of search tasks (search for local businesses and online products), and even applications that can help users complete certain tasks (language translation and math calculations).

Aggregated search systems perform two mains tasks. The first task (vertical selection) is to predict which verticals (if any) to present in response to a user's query. The second task (vertical presentation) is to predict where and how to present each selected vertical alongside the core web results.

The goal of this work is to provide a comprehensive summary of previous research in aggregated search. We first describe why aggregated search requires unique solutions. Then, we discuss different sources of evidence that are likely to be available to an aggregated search system, as well as different techniques for integrating evidence in order to make vertical selection and presentation decisions. Next, we survey different evaluation methodologies for aggregated search and discuss prior user studies that have aimed to better understand how users behave with aggregated search interfaces. Finally, we review different advanced topics in aggregated search.



J. Arguello. *Aggregated Search*. Foundations and Trends® in Information Retrieval, vol. 10, no. 5, pp. 365–502, 2016.
DOI: 10.1561/1500000052.

# 1

# Introduction

In recent years, the field of information retrieval (IR) has broadened its scope to address a wide range of information-seeking tasks. Examples include search for images, video, news, digitized books, items for sale, local businesses, scholarly articles, and even social media updates such as tweets. A common finding in empirical IR research is that different information-seeking tasks require different solutions. Specifically, different tasks require different ways of representing items in the index, different retrieval algorithms for predicting relevance, and different ways of displaying search results to users.

Different types of media may require different representations. For example, images may need to be represented using text from the surrounding context in the originating page [Feng and Lapata, 2010], social media updates may need to be represented using text obtained from the link-to URL (if one is available) [McCreadie and Macdonald, 2013], and books may need to be represented using text from an external summary page [Koolen et al., 2009]. Different search tasks may also require customized retrieval algorithms. For example, news search may require favoring recently published articles [Diaz, 2009], local business search may require favoring businesses that are geographically close [Abou-

Assaleh and Gao, 2007], and scholarly article search may require favoring articles with many citations [Lawrence et al., 1999]. Finally, different search tasks may require different ways of presenting the search results to users, by highlighting the most important attributes of the underlying item. In current systems, for example, webpage results are typically displayed using the webpage title and a summary snippet showing the context where the query terms appear on the page; items for sale are typically displayed using a thumbnail image of the product, a description, and the price; and videos are typically displayed using a stillframe of the video, a description, and the duration.

Search systems today are more diverse and specialized than ever before. In fact, search portals that aim to support different information-seeking tasks typically develop and maintain specialized search systems for different task types. Rather than attempt to address all task types with a single monolithic system, the current trend is towards a "divide and conquer" approach. Naturally, this gives rise to a new challenge: How do we provide integrated search across these widely different systems? This is the goal of *aggregated search*. The aim of aggregated search technology is to provide integrated search across a wide range of highly specialized search systems in a unified interface—a single search query box and a common presentation of results.

To date, most research in aggregated search has focused on the web search domain. For this reason, most of the research reviewed in this article will also focus on the web search domain. Commercial web search portals such as Google, Bing, and Yahoo! provide access to a wide range of specialized search services besides web search. These specialized search services are referred to as *vertical search services* or simply *verticals*. Example verticals include search engines for different types of media (e.g., images, video, news) and search services for different types of search tasks (e.g., search for local business, products for sale, scientific articles). In some cases, search portals even provide access to verticals that help users accomplish specific tasks such as language translation, unit conversation, and math calculations.

There are currently two ways that users can access vertical content. If the user wants results from a specific vertical, and if the vertical has direct search capabilities, then the user can issue the query directly to the vertical. In other cases, however, the user may not know that a vertical has relevant content, or may want results from multiple verticals at once. For this reason, an important task for commercial search providers has become the prediction and integration of relevant vertical content alongside the core web search results.

Figure 1.1 shows an example aggregated search results page (SERP) in the web domain. In response to the query "saturn", an aggregated search system decided to display news, image, and video vertical results in addition to the core web results. The most confidently relevant verticals are displayed higher on the SERP. In this case, the system predicted that the most relevant verticals were the news, images, and video verticals, respectively.

## 1.1 Aggregated Search Tasks

Most aggregated search systems follow a pipeline architecture with three subsequent sub-tasks (Figure 1.2). The first sub-task (*vertical selection*) is to predict *which* verticals (if any) are relevant to the query. One can view the vertical selection task as that of deciding which verticals should be displayed on the SERP regardless of their position. It is impractical, if not impossible, to issue the query to every available vertical. For this reason, most approaches to vertical selection base their predictions using *pre-retrieval* evidence (e.g., the query contains the term "news", the query is related to the health domain, or the query contains the name of a location).

The second sub-task (*vertical results selection*) is to predict which results from a particular vertical to present on the aggregated SERP. This sub-task has received the least attention in the research community. The vertical results selection task has a dual objective. The primary objective is to satisfy the user directly with the vertical results that are aggregated on the SERP. The secondary objective is more nuanced. Some verticals have direct search capabilities. If the user realizes that the vertical may have relevant information, he or she can navigate to the vertical, examine more vertical results, and even issue

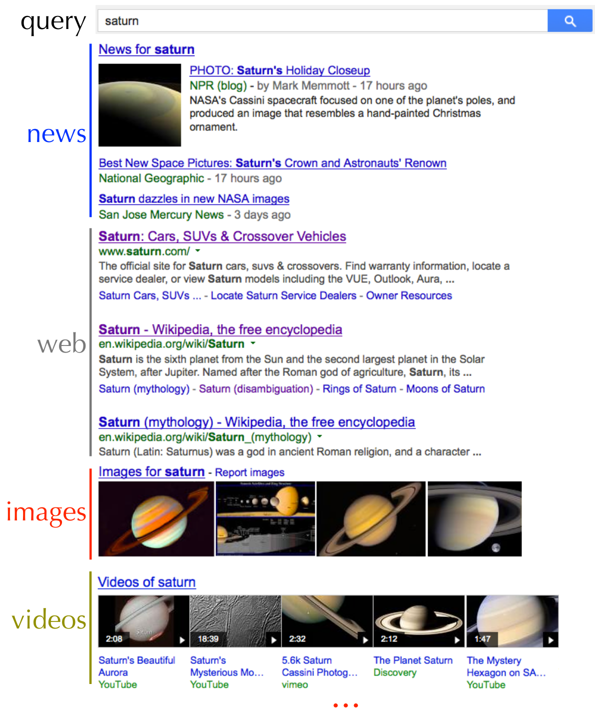


**Figure 1.1:** Aggregated SERP in the web domain (truncated). In response to the query "saturn", the aggregated search system decides to display news, image, and video vertical results in addition to the core web results. The most confidently relevant verticals are displayed higher on the SERP.

new queries to the vertical search engine. In this respect, the secondary objective of vertical results selection is to convey how the underlying vertical may have relevant content. Most aggregated search systems described in the published literature do not perform vertical results selection and simply display the top few results returned by the vertical in response to the query.

The third and final sub-task (*vertical presentation*) is to decide *where* to present each selected vertical. Different verticals are typically associated with different surrogate representations. For example, image results are displayed using thumbnails, while news results are displayed using the article title, source, publication date, and may include an op-

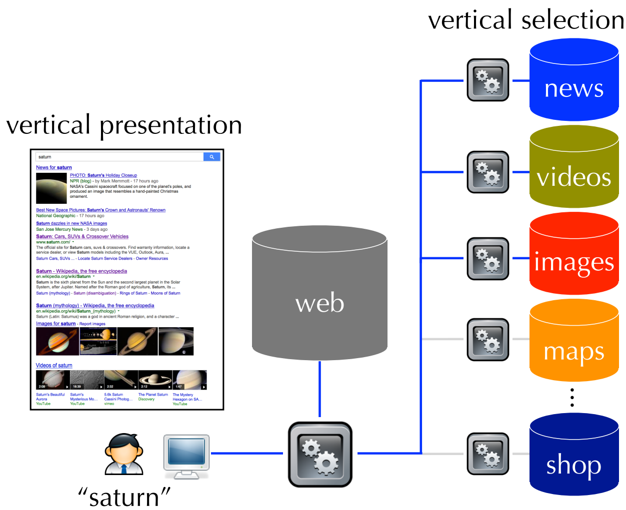


**Figure 1.2:** Aggregated search pipeline.

tional image from the underlying article. For aesthetic reasons and to better convey how the vertical may have relevant content for the current user, vertical results are typically grouped together (either stacked horizontally or vertically) on the aggregated SERP.

The goal of vertical presentation is to display the most relevant verticals in a more salient way. One common approach is to display them higher on the SERP (e.g., above the first web result). Vertical presentation happens after the query has been issued to the vertical. Thus, approaches for vertical presentation can base their predictions using pre-retrieval as well as *post-retrieval* evidence (e.g., the number of results returned by the vertical, the top retrieval scores, or the number of query-terms appearing in the top results).

## 1.2 Relation to Federated Search

While aggregated search may seem like a new technology, it is rooted in a fairly mature subfield in information retrieval known as *federated search* or *distributed information retrieval*. The goal of *federated search* is to provide integrated search across multiple collections of *textual* documents, also referred to as *resources*. Similar to aggregated search, federated search is typically decomposed into three sub-tasks.

The first sub-task (*resource representation*) is to construct a description of each distributed resource that can be used to predict which ones to search in response to a query. Approaches for resource representation differ greatly depending on whether they assume a *cooperative* or *uncooperative* environment. In a cooperative environment, resources are assumed to readily publish term statistics that can be used to model the contents of each collection [Gravano et al., 1997]. On the other hand, in an uncooperative environment, resources are assumed to only provide a search interface. In this case, resource descriptions must be constructed from sampled documents obtained via query-based sampling. In general, query-based sampling involves issuing queries to each resource and downloading results [Callan and Connell, 2001; Caverlee et al., 2006; Shokouhi et al., 2006a].

The second sub-task (*resource selection*) is to predict which resources to search in response to a query. Typically, the relevant documents are concentrated in only a few of the available resources. Resource selection approaches tend to cast the task as *resource ranking*—ranking resources based on the likelihood that they will return relevant results for the query. Existing approaches can be categorized into two types: *large document* and *small document* models. Large document models select resources based on the similarity between the query and a virtual concatenation of all the documents in the resource (or its samples). These methods treat each collection as a large document and adapt document-ranking algorithms for the purpose of ranking collections. In contrast, small document models typically proceed in two steps. First, they combine documents (or samples) from the different resources in a *centralized sample index* (CSI). Then, at query-time, they rank resources based on the top-ranked CSI results [Si and Callan, 2003a; Shokouhi, 2007; Thomas and Shokouhi, 2009].

The third sub-task (*results merging*) is to interleave the results from the different selected resources into a single ranking. Typically, this is cast as a score normalization problem [Si and Callan, 2003b]. Because different resources have different collection statistics and perhaps use different ranking algorithms, their retrieval scores may not be directly comparable. Thus, results merging requires transforming resource-*specific* scores into resource-*agnostic* scores that can be used to produce a single merged ranking. Results merging approaches typically assume that documents can be interleaved in an unconstrained fashion. The only goal is to rank the relevant documents higher on the list, irrespective of the originating resource(s).

Most federated search approaches make assumptions that do not hold true in an aggregated search environment. Thus, while there are similarities between aggregated and federated search, aggregated search requires unique solutions. Next, we discuss some of the main difference between the aggregated and federated search.

## 1.3 Differences between Aggregated and Federated Search

**Cooperative vs. uncooperative environment.** Most federated search approaches assume an uncooperative environment in which the different resources provide the system no more than the same functionality they provide their human users—a search interface. For this reason, most resource selection approaches base their predictions solely on the similarity between the input query and the documents sampled from each resource. In contrast, most aggregated search approaches assume a cooperative environment in which the different verticals are developed and maintained by the same organization. In a cooperative environment, the aggregated search system may have access to sources of evidence beyond sampled documents. For example, for verticals with direct search capabilities, alternative sources of evidence may include vertical-specific query-traffic data, click-through data, and query-reformulation data. This type of evidence conveys how users interact directly with the vertical search engine and may be helpful in predicting vertical relevance. A vertical selection system should be capable of incorporating these various sources of evidence into selection decisions.

**Heterogeneous vs. Homogeneous Content.** Most federated search approaches assume that all the distributed resources contain *textual* documents. For example, small document approaches for resource selection assume that samples from different resources can be combined in a centralized sample index (CSI), and that resources can be selected based on the top-ranked CSI results. In contrast, approaches for vertical selection need to accommodate the fact that different verticals may contains very different types of items that can not be centrally indexed and searched (e.g., news articles, images, videos, items for sale, digitized books, social media updates, etc.).

**Heterogeneous vs. Homogeneous Relevance Prediction.** Most federated search approaches apply the *same* scoring function to *every* available resource in order to predict its relevance to a query. For example, small document approaches score every resource based on the top CSI results. Similarly, large document models score every resource based on the similarity between the query and a virtual concatenation of those documents sampled from the resource. In contrast, approaches for vertical selection and presentation must be able to learn a *vertical-specific* relationship between different types of evidence and a particular vertical's relevance to a query.

To illustrate, let us consider two examples. First, certain key words are likely to predict that a particular vertical is relevant to the query. For example, the query term "news" suggests that the news vertical is relevant, while the query term "pics" suggests that the images vertical is relevant. Second, some verticals tend to be topically focused (e.g., health, auto, travel, movies). Thus, in some cases, it may be possible to predict that a particular vertical is relevant based on the general topic of the query. For example, we can predict that the health vertical is relevant to the query "swine flu" because the query is related to the health domain. Both of these examples suggest that aggregated search approaches must be able to learn a *vertical-specific* relation between certain types of evidence and the relevance of a particular vertical.

**Selection vs. Ranking.** Most federated search approaches treat resource *selection* as resource *ranking*. The goal for the system is to prioritize resources in response to a query, and to select as many or as few resource as possible given the current computational resources available. Implicit in this formulation of the resource selection task is the assumption that *exhaustive* search produces a good retrieval and that the goal for the system is to approximate this retrieval by selecting only a few resources. In contrast, vertical selection requires predicting which verticals are relevant to the query and which verticals are not. In some cases, the system may decide that none of the available verticals are relevant. Thus, vertical selection requires approaches that can make binary predictions for each candidate resource.

**Constrained vs. Unconstrained Results Presentation.** Finally, most federated search approaches assume that the results from the different selected resources can be interleaved in an unconstrained fashion. In contrast, most aggregated search approaches assume that the results from the same vertical must be presented together on the SERP in the form of a vertical block. This is mostly done for aesthetic reasons and to provide an easy-to-parse overview of how the vertical may have relevant content for the query. Vertical presentation approaches must address the unique challenge of deciding where to present each selected vertical on the SERP.

## 1.4 Overview of Aggregated Search Algorithms

Most successful approaches for vertical selection and presentation use machine learning to combine a wide range of evidence as input features to the model. Features can be generated from the query, from the vertical, or from the query-vertical pair. For example, a type of query feature might consider whether the query contains the keyword "news", a type of vertical feature might consider the number of recent clicks on the vertical results, and a type of query-vertical might estimate the number of query-related documents in the underlying vertical collection. The most effective approaches for vertical selection and presentation make creative use of the different sources of evidence available to the system, including vertical-specific query-log data, sampled vertical documents, and previous user interactions with vertical content.

While evidence integration is key to aggregated search, it also poses two main challenges. The first challenge is that not all features may be available for all verticals. For example, some verticals cannot be directly searched by users. Consider the weather vertical in most commercial search portals. Users cannot typically go directly to the weather vertical and issue a query. Thus, features generated from the vertical query-log will not be available for verticals that are not directly searchable. Similarly, some verticals are not associated with an underlying collection of documents. Consider the calculator, language translation, and finance verticals in most commercial search portals. Features that consider the similarity between the query and the documents in the underlying vertical will not be available for such verticals. In this respect, approaches for vertical selection and presentation must deal with the fact that different verticals may require different feature representations.

The second challenge is that, even if a feature is available for all verticals, it may not be *equally* predictive across verticals. For example, certain verticals are clicked more than others. For example, a news vertical is likely to have more clicks than a weather vertical, which is designed to display the necessary information directly on the SERP. Features derived from click data (e.g., the number of recent clicks on the vertical results) may be more predictive for verticals that have more clicks. Alternatively, a feature may be *positively* predictive for one vertical and *negative* predictive for another. Consider, for example, a feature that measures whether the query is related to the travel domain. This feature is likely to be positively predictive for a travel-related vertical, but negatively predictive for a vertical that focuses on a different domain. In this respect, approaches for vertical selection and presentation must deal with the fact that different verticals may require learning a *vertical-specific* relationship between certain features and a vertical's relevance.

Given the two challenges outlined above, approaches for vertical selection typically learn a different model for each candidate vertical. In this way, each model can adopt a different feature representation and can learn a vertical-specific relationship between feature values and the relevance of the particular vertical. Vertical presentation requires resolving contention between different verticals to be displayed on the SERP. Put differently, vertical presentation requires predicting the degree of relevance of a vertical relative to the web results and relative to other verticals to be displayed. Approaches for vertical presentation can be categorized into two types: pointwise and pairwise interleaving methods. Pointwise methods learn to predict the *degree* of relevance of each vertical block or module in response to a query. Vertical blocks are positioned according to their predicted relevance to the query. Pairwise methods learn to predict the relative relevance between *pairs* of vertical and/or web blocks or modules. Vertical blocks are positioned such that they are maximally consistent with the pairwise preferences predicted by the system.

## 1.5 Related Topics

In this review, we focus on aggregated search in the web domain, where systems combine results from heterogeneous sources (or verticals) into a single presentation. We cover a wide range of topics, including prediction, evaluation, and studies of user behavior.

We focus on the web domain because most of the published research has been done in this domain. However, the task of searching and integrating information from heterogeneous sources happens in other domains within the broad field of information retrieval. For example, in desktop search, the system needs to search across different types of files, which may require different indexing structures, ranking algorithms, and ways of presenting the search results. Similarly, news aggregators are responsible for combining content from different input streams, such as news articles, images, videos, and social media updates.

In this section, we describe related areas of IR research that may benefit from the algorithms, evaluation methods, and studies described in this review.

### 1.5.1 Full-text Search in Peer-to-Peer Networks

A peer-to-peer (P2P) network is defined as a network of independent computing resources that do not require a centralized authority to coordinate and perform tasks. A *hierarchical* (P2P) network is one with three types of peers: (1) peers that provide search for a particular collection, such as a digital library (*providers*), (2) peers that originate information requests for the network (*consumers*), and (3) peers that propagate information requests to neighboring peers and send results back to the corresponding consumer (*hubs*). Hubs perform the three main tasks associated with aggregated search: (1) representing the contents of neighboring peers (i.e., direct providers and other hubs), (2) sending information requests to the neighboring peers most likely to deliver relevant content, and (3) merging the results returned by the selected peers and sending these back to the appropriate consumer. Lu [2007] proposed several approaches for these three different tasks that build upon traditional federated search techniques (where there is a centralized federated search system that has direct access to all available resources).

The techniques discussed in this review might be useful for the tasks of query routing and results merging in P2P networks that provide distributed search capabilities. Beverly and Afergan [2007], for example, proposed a machine learning, evidence integration approach for neighbor selection in P2P networks.

### 1.5.2 Desktop Search

The goal of *desktop search* is to facilitate search over files stored in a user's desktop computer. One of the main challenges in desktop search is that different file types are associated very different field structures and meta-data. Kim and Croft [2010] developed and evaluated a desktop search system that maintains different indexes for different file types. Given a query, the proposed system performs the three basic steps associated with aggregated search: file-type prediction, file-type-specific ranking, and results merging. Much like the vertical selection methods covered in this review, the proposed file-type prediction approach combined multiple types of evidence as features for a machine learned model, for example, the similarity between the query and document meta-data, the similarity between the query and previously run queries with clicks on a particular file-type, and the presence of certain query keywords such as "email" or "pdf". As one might expect, the evidence integration approach to file-type prediction outperformed the best approach using a single source of evidence.

### 1.5.3 Selective Search

The aim of *selective search* is to enable efficient and effective search from large text collections in environments with modest computational resources [Kulkarni and Callan, 2015]. First, the system partitions the large text collection into smaller *topical* sub-collections or *shards*. Then, in response to a query, the system predicts which few shards are most likely to have relevant documents and merges their results. Selective search is highly motivated by the *cluster hypothesis*, which states that similar documents (ideally assigned to the same shard) tend to be relevant to same information needs [van Rijsbergen, 1979]. Shard representation and selection can be performed using existing federated search techniques, and results merging is relatively straightforward because the system has access to global term statistics can be used to compute comparable retrieval scores. The critical step in selective search is partitioning the collection into topical shards. Kulkarni and Callan [2015] proposed a variant of the well-known K-means clustering algorithm that operates on a sample of documents from the collection. Experimental results show that selective search can greatly reduce computational costs and latency, and can yield retrieval performance comparable to exhaustive search, particularly for precision-oriented tasks.

While current shard-selection techniques do not combine multiple types of evidence to make predictions, prior work on text-based federated search used machine learning to combine a wide range of features for the task of resource selection [Arguello et al., 2009a; Hong et al., 2010]. In particular, because shards are topically focused, the query category features discussed later in Section 2.3 might contribute valuable evidence for shard selection.

### 1.5.4 Contextual Suggestion

The goal of *contextual suggestion* is to recommend points-of-interest (POIs) to a user in a particular context (i.e., in a particular location, at a particular time) [Dean-Hall et al., 2012, 2013, 2014, 2015]. The system is assumed to have access to ratings on previously recommended POIs for the same user in different contexts.

Zhuang et al. [2011] describe a mobile contextual suggestion system with an aggregated search architecture. Rather than index and retrieve all POIs using a single system, the proposed approach is to build different indexes and rankers for different POI-types (e.g., restaurants, coffee shops, bars, tourist attractions, etc.) The system recommends POIs to a user in a particular context in two steps. First, the system predicts the appropriateness of a particular POI-type for the given context, and then it ranks POIs of a particular type if the user requests to see those results. Similar to aggregated search, the proposed architecture has two main benefits. First, the system can use different models for predicting relevance for each POI-type. For example, the system can learn that restaurants are more relevant during meal times and that bars are more relevant in the evening. Second, the system can learn different rankers for different POI-types. For example, the system can determine that close proximity to the user is more important for coffee shops than for tourist attractions (assuming users are more willing to travel longer distances for the latter).

### 1.5.5 Search Across Heterogeneous Social Networks

In certain cases, a user may belong to multiple social networks and may want to receive updates from different networks in a unified interface. Bian et al. [2012] proposed an algorithm for ranking social network updates originating from different networks. The main challenge is that different networks may be associated with different sources of evidence that can be used to predict the relevance of an update for a particular user. Consider a user who wants to receive aggregated updates from both Facebook and Twitter. Some sources of evidence are common to both networks (e.g., Does the update contain a URL?). However, other features may aim to exploit the same type of evidence, but be associated with very different numerical ranges across networks (e.g., number of comments on Facebook and number of retweets on Twitter). Moreover, some features may only exist in one network and not the other (e.g., the number of Facebook chat messages between the user and the author of an update). Rather than rank candidate updates from different networks using a single model, Bian et al. [2012] describe a "divide and conquer" approach that learns network-specific rankers and combines their output rankings into a single merged list.

Lee et al. [2012] focused on the task of ranking social media updates and used two test collections: one generated from Facebook updates and another generated from Twitter updates. The authors did not attempt the task of constructing a single, merged ranking. However, the authors concluded that combining updates from different heterogeneous social networks into a single ranked list is an interesting research direction for future work.

### 1.5.6 News Aggregators

News content aggregators such as the Yahoo! homepage or the New York Times homepage combine results from different heterogeneous data streams into a single presentation. Data streams may include news articles from different sources, images, videos, audio interviews, blog posts, and social media updates such as tweets. The system is responsible for predicting which items to display from each data stream and where [Bharat et al., 1998; Krakovsky, 2011]. Different data streams are likely to be associated with very different types of evidence that can be used to predict relevance. Thus, news aggregators are likely to benefit from a "divide and conquer" approach—building customized rankers for different data streams and a system that predicts which content to display and where.

One interesting aspect of news aggregation is that in some cases, the system may want to show results from different data streams that are related to the same topic. For example, the system may want to display news, images, videos, and opinionated tweets about the same trending news story. Hong et al. [2011] proposed an approach for finding related content in different data streams. In the context of aggregated search, the results from different sources aggregated on the search results page are typically independent of each other. However, identifying related results in different sources or verticals may be an interesting direction for future work.

## 1.6 Related Surveys

As mentioned above, aggregated search is related to the subfield of federated search or distributed information retrieval, where the goal is to provide integrated search across multiple *textual* collections. Shokouhi and Si [2011] provide an extensive review of the state of the art in federated search, and review methods for all three federated search sub-tasks: resource representation, selection, and results merging.

Chapter 4 in this review focuses on methods of aggregated search evaluation. Online evaluation approaches learn about a system's performance from user interactions in a live environment. In the context of aggregated search, vertical selection approaches can be evaluated by considering user's clicks on the vertical results. Interpreting user interactions with a SERP is complicated by the fact that users are influenced by factors other than relevance, such as position and visual salience. Hofmann et al. [2016] provide an extensive survey of approaches for online evaluation using real users.

The current survey is most closely related to the book chapter titled "Aggregated Vertical Search" appearing in Long and Chang [2014]. However, the current survey is different in several respects. First, it includes new solutions, evaluation methods, and user studies published since 2014. In recent years, studies have proposed and tested new evaluation metrics for aggregated search [Zhou et al., 2013b]. Furthermore, recent studies have investigated different factors that may affect search behavior and performance with aggregated search interfaces. For example, recent work investigates how users visually scan an aggregated SERP [Liu et al., 2015], how the results from one source on the SERP can influence user engagement with the results from other sources [Arguello and Capra, 2016; Bota et al., 2016], and how users' cognitive abilities can affect different search behaviors and outcomes [Turpin et al., 2016].

Furthermore, this review covers more special topics in aggregated search. For example, it surveys recent work on *composite retrieval*, where the goal for the system is to combine results from different sources, but to organize them by how they satisfy different *aspects* of the user's task. Also, it covers recent work on aggregated search for children, who exhibit different search behaviors than adults and require unique aggregated search solutions [Duarte-Torres and Weber, 2011].

## 1.7 Outline

As previously mentioned, the most effective approaches for vertical selection and presentation use machine learning to combine different types of evidence as features. Chapter 2 reviews different features used in prior work. These include features that derive evidence from vertical content, from queries issued directly to the vertical by users, and from previous users' interactions with the results from a particular vertical.

In a sense, vertical selection and presentation have a common goal—to predict the degree of relevance of a vertical to a user's query. In Chapter 2, we remain somewhat agnostic as to whether a particular feature is more appropriate for one task versus the other. That said, certain features (referred to as *post-retrieval* features) require issuing the query to the candidate vertical. Thus, in some places, we emphasize that post-retrieval features may be more appropriate for vertical presentation.

Chapter 3 focuses on evidence combination approaches for vertical selection and presentation. The main challenge in vertical selection and presentation is that certain features may be predictive for one vertical, but not another. For example, the publication age of the top vertical results may be predictive for the news vertical, but not the image vertical. Moreover, certain features may be *positively* predictive for one vertical, but *negatively* predictive for another. For example, the query term "news" is positively predictive for the news vertical, but negatively predictive for the image vertical. For this reason, in Chapter 3 we focus on approaches that can exploit a *vertical-specific* relationship between different features and the relevance of a particular vertical.

Chapter 4 focuses on evaluation methodologies and metrics for aggregated search. Evaluation is a critical component of all information retrieval techniques and a research area in its own right. We start with vertical selection and then cover end-to-end evaluation, which includes selection and presentation. We cover evaluation methodologies based on re-usable test collections, which typically include a set of evaluation queries, cached results from the different sources, and human-produced relevance judgements. We also discuss on-line evaluation methodologies based on implicit feedback from real users in an operational setting.

Chapter 5 reviews user studies aimed at further understanding what users want from an aggregated search system and how they behave. We cover studies where the goal is to determine the extent to which a particular evaluation metric correlates with user satisfaction, and studies where the goal is to understand how different characteristics of the interface, the search task, and the user can affect outcome measures associated with the user's perceptions about the system and their performance.

Chapter 6 reviews special topics in aggregated search. Here, we touch upon algorithms for predicting how a user will visually scan a particular aggregated SERP, methods for leveraging implicit feedback in order to improve performance, and approaches for learning a model for a new vertical with little human-produced training data. Furthermore, we review the new task of *composite retrieval*, where the goal is to organize results from different sources based on different *aspects* associated with the task. Finally, we discuss aggregated search for children, who exhibit different behavior than adults and require unique solutions.

Finally, in Chapter 7, we conclude by highlighting the main trends in aggregated search and discussing short-term and long-term areas for future work.

# References

Abou-Assaleh, T. and Gao, W. Geographic ranking for a local search engine. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 911–911, New York, NY, USA, 2007.

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. Diversifying search results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA, 2009.

Al-Maskari, A. and Sanderson, M. The effect of user characteristics on search effectiveness in information retrieval. *Information Processing and Management*, 47(5):719–729, September 2011.

Arguello, J. Improving aggregated search coherence. In *Proceedings of the 37th European Conference on Advances in Information Retrieval*, volume 9022 of *ECIR '15*, pages 25–36, Springer-Verlag, Berlin, Heidelberg, 2015.

Arguello, J. and Capra, R. The effects of aggregated search coherence on search behavior. *ACM Transactions on Information Systems*, 11(1), 2016.

Arguello, J. and Capra, R. The effect of aggregated search coherence on search behavior. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1293–1302, New York, NY, USA, 2012.

Arguello, J. and Capra, R. The effects of vertical rank and border on aggregated search coherence and search behavior. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 539–548, New York, NY, USA, 2014.

Arguello, J., Callan, J., and Diaz, F. Classification-based resource selection. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1277–1286, New York, NY, USA, 2009a.

Arguello, J., Diaz, F., Callan, J., and Crespo, J.-F. Sources of evidence for vertical selection. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 315–322, New York, NY, USA, 2009b.

Arguello, J., Diaz, F., and Paiement, J.-F. Vertical selection in the presence of unlabeled verticals. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 691–698, New York, NY, USA, 2010.

Arguello, J., Diaz, F., and Callan, J. Learning to aggregate vertical results into web search results. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 201–210, New York, NY, USA, 2011a.

Arguello, J., Diaz, F., Callan, J., and Carterette, B. A methodology for evaluating aggregated search results. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR '11, pages 141–152, Springer-Verlag, Berlin, Heidelberg, 2011b.

Arguello, J., Wu, W.-C., Kelly, D., and Edwards, A. Task complexity, vertical display, and user interaction in aggregated search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 435–444, New York, NY, USA, 2012.

Arguello, J., Capra, R., and Wu, W.-C. Factors affecting aggregated search coherence and search behavior. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM '13, pages 1989–1998, New York, NY, USA, 2013.

Aula, A., Khan, R. M., and Guan, Z. How does search behavior change as search becomes more difficult? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 35–44, New York, NY, USA, 2010.

Bailey, P., Craswell, N., White, R. W., Chen, L., Satyanarayana, A., and Tahaghoghi, S. M. Evaluating search systems using result page context. In *Proceedings of the 3rd Symposium on Information Interaction in Context*, IIiX '10, pages 105–114, New York, NY, USA, 2010a.

Bailey, P., Craswell, N., White, R. W., Chen, L., Satyanarayana, A., and Tahaghoghi, S. Evaluating whole-page relevance. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 767–768, New York, NY, USA, 2010b.

Bennett, P. N., Svore, K., and Dumais, S. T. Classification-enhanced ranking. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 111–120, New York, NY, USA, 2010.

Beverly, R. and Afergan, M. Machine learning for efficient neighbor selection in unstructured p2p networks. In *Proceedings of the 2Nd USENIX Workshop on Tackling Computer Systems Problems with Machine Learning Techniques*, SYSML '07, pages 1:1–1:6, USENIX Association, Berkeley, CA, USA, 2007.

Bharat, K., Kamba, T., and Albers, M. Personalized, interactive news on the web. *Multimedia Systems*, 6(5):349–358, 1998.

Bian, J., Chang, Y., Fu, Y., and Chen, W.-Y. Learning to blend vitality rankings from heterogeneous social networks. *Neurocomputing*, 97:390–397, 2012. ISSN 0925-2312.

Bilal, D. Children's use of the yahooligans! web search engine. iii. cognitive and physical behaviors on fully self-generated search tasks. *Journal of the American Society for Information Science and Technology*, 53(13):1170–1183, 2002.

Bota, H., Zhou, K., Jose, J. M., and Lalmas, M. Composite retrieval of heterogeneous web search. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 119–130, New York, NY, USA, 2014.

Bota, H., Zhou, K., and Jose, J. J. *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, chapter Exploring Composite Retrieval from the Users' Perspective, pages 13–24. Springer International Publishing, 2015.

Bota, H., Zhou, K., and Jose, J. M. Playing your cards right: The effect of entity cards on search behaviour and workload. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 131–140, New York, NY, USA, 2016.

Brennan, K., Kelly, D., and Arguello, J. The effect of cognitive abilities on information search for tasks of varying levels of complexity. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 165–174, New York, NY, USA, 2014.

Broder, A. A taxonomy of web search. *SIGIR Forum*, 36(2), September 2002.

Broder, A., Fontoura, M., Josifovski, V., and Riedel, L. A semantic approach to contextual advertising. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 559–566, New York, NY, USA, 2007.

Bron, M., van Gorp, J., Nack, F., Baltussen, L. B., and de Rijke, M. Aggregated search interface preferences in multi-session search tasks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 123–132, New York, NY, USA, 2013.

Callan, J. and Connell, M. Query-based sampling of text databases. *ACM Transactions on Information Systems*, 19:97–130, 2001.

Capra, R., Arguello, J., and Scholer, F. Augmenting web search surrogates with images. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM '13, pages 399–408, New York, NY, USA, 2013.

Carterette, B., Kanoulas, E., Hall, M., and Clough, P. Overview of the trec 2014 session track. In *Proceedings of the 24th Text Retrieval Conference*, TREC '14, NIST. 2014.

Caverlee, J., Liu, L., and Bae, J. Distributed query sampling: A quality-conscious approach. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 340–347, New York, NY, USA, 2006.

Chapelle, O., Joachims, T., Radlinski, F., and Yue, Y. Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions of Information Systems*, 30(1):6:1–6:41, 2012.

Chen, D., Chen, W., Wang, H., Chen, Z., and Yang, Q. Beyond ten blue links: Enabling user click modeling in federated web search. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 463–472, New York, NY, USA, 2012.

Chen, K., Bai, J., and Zheng, Z. Ranking function adaptation with boosting trees. *ACM Transactions of Information Systems*, 29(4):18:1–18:31, December 2011.

Chen, Y., Liu, Y., Zhou, K., Wang, M., Zhang, M., and Ma, S. Does vertical bring more satisfaction?: Predicting search satisfaction in a heterogeneous environment. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1581–1590, New York, NY, USA, 2015.

ChildTrends. Home computer access and internet use. http://www.childtrends.org/?indicators=home-computer-access, 2015. Accessed: 2016-05-31.

Chuklin, A., Schuth, A., Hofmann, K., Serdyukov, P., and de Rijke, M. Evaluating aggregated search using interleaving. In *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '13, pages 669–678, New York, NY, USA, 2013a.

Chuklin, A., Serdyukov, P., and de Rijke, M. Click model-based information retrieval metrics. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 493–502, New York, NY, USA, 2013b.

Cleverdon, C. W. The aslib cranfield research project on the comparative efficiency of indexing systems. *Aslib Proceedings*, 12(12):421–431, 1960.

Comscore. Digital Future in Focus U.S. 2015. Technical report, 2015.

Cronen-Townsend, S., Zhou, Y., and Croft, W. B. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 299–306, New York, NY, USA, 2002.

Dean-Hall, A., Clarke, C. L. A., Kamps, J., Thomas, P., and Voorhees, E. Overview of the trec 2012 contextual suggestion track. In *Proceedings of the 21st Text Retrieval Conference*, TREC '12, NIST. 2012.

Dean-Hall, A., Clarke, C. L. A., Simone, N., Kamps, J., Thomas, P., and Voorhees, E. Overview of the trec 2013 contextual suggestion track. In *Proceedings of the 22nd Text Retrieval Conference*, TREC '13, NIST. 2013.

Dean-Hall, A., Clarke, C. L. A., Kamps, J., Thomas, P., and Voorhees, E. Overview of the trec 2014 contextual suggestion track. In *Proceedings of the 23rd Text Retrieval Conference*, TREC '14, NIST. 2014.

Dean-Hall, A., Clarke, C. L. A., Kamps, J., Kiseleva, J., and Voorhees, E. Overview of the trec 2014 contextual suggestion track. In *Proceedings of the 24th Text Retrieval Conference*, TREC '15, NIST. 2015.

Demeester, T., Trieschnigg, D., Nguyen, D., and Hiemstra, D. Overview of the trec 2013 federated web search track. In *Proceedings of the 23rd Text Retrieval Conference*, TREC '13, NIST. 2013.

Demeester, T., Trieschnigg, D., Nguyen, D., Hiemstra, D., and Zhou, K. Overview of the trec 2014 federated web search track. In *Proceedings of the 23rd Text Retrieval Conference*, TREC '14, NIST. 2014.

Diaz, F. Performance prediction using spatial autocorrelation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 583–590, New York, NY, USA, 2007.

Diaz, F. Integration of news content into web results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 182–191, New York, NY, USA, 2009.

Diaz, F. and Arguello, J. Adaptation of offline vertical selection predictions in the presence of user feedback. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 323–330, New York, NY, USA, 2009.

Diaz, F., White, R., Buscher, G., and Liebling, D. Robust models of mouse movement on dynamic web search results pages. In *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '13, pages 1451–1460, New York, NY, USA, 2013.

Duarte-Torres, S. and Weber, I. What and how children search on the web. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 393–402, New York, NY, USA, 2011.

Duarte-Torres, S., Hiemstra, D., and Serdyukov, P. Query log analysis in the context of information retrieval for children. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 847–848, New York, NY, USA, 2010.

Duarte-Torres, S., Hiemstra, D., and Huibers, T. Vertical selection in the information domain of children. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, pages 57–66, New York, NY, USA, 2013.

Duarte-Torres, S., Weber, I., and Hiemstra, D. Analysis of search and browsing behavior of young users on the web. *ACM Transactions on the Web*, (2):7:1–7:54, 2014.

Dumais, S., Cutrell, E., and Chen, H. Optimizing search by showing results in context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 277–284, New York, NY, USA, 2001.

Dworman, G. and Rosenbaum, S. Helping users to use help: Improving interaction with help systems. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, pages 1717–1718, New York, NY, USA, 2004.

Ekstrom, R., French, J., Harman, H., and Dermen, D. *Kit of Factor-Referenced Cognitive Tests*. Educational Testing Service, Princeton, NJ, USA, 1979.

Feng, Y. and Lapata, M. Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 831–839, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010.

Fleiss, J. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

Friedman, J. H. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378, February 2002.

Gravano, L., Chang, C.-C. K., García-Molina, H., and Paepcke, A. Starts: Stanford proposal for internet meta-searching. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, SIGMOD '97, pages 207–218, New York, NY, USA, 1997.

Guy, I. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 35–44, New York, NY, USA, 2016.

Hassan, A., White, R. W., Dumais, S. T., and Wang, Y.-M. Struggling or exploring?: Disambiguating long search sessions. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 53–62, New York, NY, USA, 2014.

2010. Hauff, C. *Predicting the Effectiveness of Queries and Retrieval Systems.* dissertation, Univeristy of Twente, 2010.

Hofmann, K., Whiteson, S., and de Rijke, M. A probabilistic method for inferring preferences from clicks. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 249–258, New York, NY, USA, 2011.

Hofmann, K., Behr, F., and Radlinski, F. On caption bias in interleaving experiments. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 115–124, New York, NY, USA, 2012.

Hofmann, K., Whiteson, S., and Rijke, M. D. Fidelity, soundness, and efficiency of interleaved comparison methods. *ACM Transactions of Information Systems*, 31(4):1–43, November 2013.

Hofmann, K., Li, L. L., and Radlinski, F. Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval*, 10:1–117, June 2016.

Hong, D., Si, L., Bracke, P., Witt, M., and Juchcinski, T. A joint probabilistic classification model for resource selection. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 98–105, New York, NY, USA, 2010.

Hong, L., Dom, B., Gurumurthy, S., and Tsioutsiouliklis, K. A time-dependent topic model for multiple text streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 832–840, New York, NY, USA, 2011.

Huang, J., White, R. W., and Dumais, S. No clicks, no problem: Using cursor movements to understand and improve search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1225–1234, New York, NY, USA, 2011.

Huang, J., White, R., and Buscher, G. User see, user point: Gaze and cursor alignment in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1341–1350, New York, NY, USA, 2012a.

Huang, J., White, R. W., Buscher, G., and Wang, K. Improving searcher models using mouse cursor activity. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 195–204, New York, NY, USA, 2012b.

Jansen, B. J., Booth, D. L., and Spink, A. Determining the informational, navigational, and transactional intent of web queries. *Information Processing and Management*, 44(3):1251–1266, May 2008.

Jansen, B. J., Booth, D., and Smith, B. Using the taxonomy of cognitive learning to model online searching. *Information Processing and Management*, 45(6):643–663, November 2009.

Järvelin, K. and Kekäläinen, J. Cumulated gain-based evaluation of ir techniques. *ACM Transactions of Informaton Systems*, 20(4):422–446, 2002.

Jeon, J., Croft, W. B., Lee, J. H., and Park, S. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 228–235, New York, NY, USA, 2006.

Jiang, J., Hassan Awadallah, A., Jones, R., Ozertem, U., Zitouni, I., Gurunath Kulkarni, R., and Khan, O. Z. Automatic online evaluation of intelligent assistants. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, New York, NY, USA, 2015.

Jie, L., Lamkhede, S., Sapra, R., Hsu, E., Song, H., and Chang, Y. A unified search federation system based on online user feedback. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1195–1203, New York, NY, USA, 2013.

Joachims, T. Optimizing search engines using clickthrough data. In *Proceedings of the 8thh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA, 2002.

Khelghati, M., Hiemstra, D., and van Keulen, M. Size estimation of non-cooperative data collections. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications and Services*, IIWAS '12, pages 239–246, New York, NY, USA, 2012.

Kim, J. and Croft, W. B. Ranking using multiple document types in desktop search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 50–57, New York, NY, USA, 2010.

Kiseleva, J., Williams, K., Hassan Awadallah, A., Crook, A. C., Zitouni, I., and Anastasakos, T. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 45–54, New York, NY, USA, 2016a.

Kiseleva, J., Williams, K., Jiang, J., Hassan Awadallah, A., Crook, A. C., Zitouni, I., and Anastasakos, T. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 121–130, New York, NY, USA, 2016b.

Koffka, K. *Principles of Gestalt psychology.* Harcourt, New York, 1935.

König, A. C., Gamon, M., and Wu, Q. Click-through prediction for news queries. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 347–354, New York, NY, USA, 2009.

Koolen, M., Kazai, G., and Craswell, N. Wikipedia pages as entry points for book search. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 44–53, New York, NY, USA, 2009.

Krakovsky, M. All the news that's fit for you. *Communications of the ACM*, 54(6):20–21, 2011.

Kulkarni, A. and Callan, J. Selective search: Efficient and effective search of large textual collections. *ACM Transactions on Information Systems*, 33 (4):1–33, 2015.

Kumar, R. and Vassilvitskii, S. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 571–580, New York, NY, USA, 2010.

Lagun, D. and Agichtein, E. Inferring searcher attention by jointly modeling user interactions and content salience. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 483–492, New York, NY, USA, 2015.

Lagun, D., Ageev, M., Guo, Q., and Agichtein, E. Discovering common motifs in cursor movement data for improving web search. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 183–192, New York, NY, USA, 2014a.

Lagun, D., Hsieh, C.-H., Webster, D., and Navalpakkam, V. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 113–122, New York, NY, USA, 2014b.

Lavrenko, V. and Croft, W. B. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 120–127, New York, NY, USA, 2001.

Lawrence, S., Bollacker, K., and Giles, C. L. Indexing and retrieval of scientific literature. In *Proceedings of the 8th International Conference on Information and Knowledge Management*, CIKM '99, pages 139–146, New York, NY, USA, 1999.

Lee, C.-J., Croft, W. B., and Kim, J. Y. Evaluating search in personal social media collections. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 683–692, New York, NY, USA, 2012.

Li, J., Huffman, S., and Tokuda, A. Good abandonment in mobile and pc internet search. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 43–50, New York, NY, USA, 2009.

Li, X., Wang, Y.-Y., and Acero, A. Learning query intent from regularized click graphs. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 339–346, New York, NY, USA, 2008.

Liu, J., Cole, M. J., Liu, C., Bierig, R., Gwizdka, J., Belkin, N. J., Zhang, J., and Zhang, X. Search behaviors in different task types. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 69–78, New York, NY, USA, 2010a.

Liu, J., Liu, C., Gwizdka, J., and Belkin, N. J. Can search systems detect users' task difficulty?: Some behavioral signals. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 845–846, New York, NY, USA, 2010b.

Liu, J., Liu, C., Cole, M., Belkin, N. J., and Zhang, X. Exploring and predicting search task difficulty. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1313–1322, New York, NY, USA, 2012.

Liu, K.-L., Santoso, A., Yu, C., and Meng, W. Discovering the representative of a search engine. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 577–579, New York, NY, USA, 2001.

Liu, K.-L., Yu, C., and Meng, W. Discovering the representative of a search engine. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, pages 652–654, New York, NY, USA, 2002.

Liu, Y., Liu, Z., Zhou, K., Wang, M., Luan, H., Wang, C., Zhang, M., and Ma, S. Predicting search user examination with visual saliency. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 619–628, New York, NY, USA, 2016.

Liu, Z., Liu, Y., Zhou, K., Zhang, M., and Ma, S. Influence of vertical result in web search examination. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 193–202, New York, NY, USA, 2015.

Long, B. and Chang, Y. *Relevance Ranking for Vertical Search Engines*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2014.

2007. Lu, J. *Full-text Federated Search in Peer-to-peer Networks*. PhD thesis, Pittsburgh, PA, USA, 2007.

Luo, C., Liu, Y., Zhang, M., and Ma, S. *Query Ambiguity Identification Based on User Behavior Information*, pages 36–47. AIRS 2014. Springer International Publishing, 2014.

Lv, Y., Moon, T., Kolari, P., Zheng, Z., Wang, X., and Chang, Y. Learning to model relatedness for news recommendation. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 57–66, New York, NY, USA, 2011.

Markov, I., Kharitonov, E., Nikulin, V., Serdyukov, P., de Rijke, M., and Crestani, F. Vertical-aware click model-based effectiveness metrics. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1867–1870, New York, NY, USA, 2014.

McCreadie, R. and Macdonald, C. Relevance in microblogs: Enhancing tweet retrieval using hyperlinked documents. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 189–196, Paris, France, France, 2013.

Metzler, D., Dumais, S., and Meek, C. Similarity measures for short segments of text. In *Proceedings of the 29th European Conference on IR Research*, ECIR'07, pages 16–27, Springer-Verlag, Berlin, Heidelberg, 2007.

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

Nanba, H., Sakai, T., Kando, N., ana Koji Eguchi, A. K., Hatano, K., Shimizu, T., Hirate, Y., and Fujii, A. Nexti at ntcir-12 imine-2 task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, NTCIR '16, National Institute of Informatics. 2016.

Navalpakkam, V., Jentzsch, L., Sayres, R., Ravi, S., Ahmed, A., and Smola, A. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 953–964, New York, NY, USA, 2013.

Nygren, E. Between the clicks: Skilled users scanning of pages. In *Proceedings of Designing for the Web: Empirical Studies*, 1996.

Palmer, S. E. Common region: A new principle of perceptual grouping. *Cognitive Psychology*, 24(3):436 – 447, 1992.

Ponnuswami, A. K., Pattabiraman, K., Brand, D., and Kanungo, T. Model characterization curves for federated search using click-logs: Predicting user engagement metrics for the span of feasible operating points. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 67–76, New York, NY, USA, 2011a.

Ponnuswami, A. K., Pattabiraman, K., Wu, Q., Gilad-Bachrach, R., and Kanungo, T. On composition of a federated web search result page: Using online users to provide pairwise preference for heterogeneous verticals. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 715–724, New York, NY, USA, 2011b.

Radlinski, F. and Joachims, T. Query chains: Learning to rank from implicit feedback. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 239–248, New York, NY, USA, 2005.

Radlinski, F., Kurup, M., and Joachims, T. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 43–52, New York, NY, USA, 2008.

Sahami, M. and Heilman, T. D. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 377–386, New York, NY, USA, 2006.

Sanderson, M. Ambiguous queries: Test collections need more sense. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 499–506, New York, NY, USA, 2008.

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

Santos, R. L. T., Macdonald, C., and Ounis, I. Aggregated search result diversification. In *Proceedings of the 3rd International Conference on Advances in Information Retrieval Theory*, ICTIR '11, pages 250–261, Springer-Verlag, Berlin, Heidelberg, 2011.

Schulze, M. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Social Choice and Welfare*, 36(2):267–303, 2011.

Seo, J., Croft, W. B., Kim, K. H., and Lee, J. H. Smoothing click counts for aggregated vertical search. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 387–398, Springer-Verlag, Berlin, Heidelberg, 2011.

Shen, D., Pan, R., Sun, J.-T., Pan, J. J., Wu, K., Yin, J., and Yang, Q. Query enrichment for web-query classification. *ACM Transactions of Information Systems*, 24(3):320–352, 2006.

Shokouhi, M. Central-rank-based collection selection in uncooperative distributed information retrieval. In *Proceedings of the 29th European Conference on IR Research*, ECIR'07, pages 160–172, Springer-Verlag, Berlin, Heidelberg, 2007.

Shokouhi, M. Learning to personalize query auto-completion. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 103–112, New York, NY, USA, 2013.

Shokouhi, M. and Si, L. Federated search. *Foundations and Trends in Information Retrieval*, 5(1):1–102, January 2011.

Shokouhi, M., Scholer, F., and Zobel, J. Sample sizes for query probing in uncooperative distributed information retrieval. In *Proceedings of the 8th Asia-Pacific Web Conference on Frontiers of WWW Research and Development*, APWeb'06, pages 63–75, Springer-Verlag, Berlin, Heidelberg, 2006a.

Shokouhi, M., Zobel, J., Scholer, F., and Tahaghoghi, S. M. M. Capturing collection size for distributed non-cooperative retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 316–323, New York, NY, USA, 2006b.

Si, L. and Callan, J. Relevant document distribution estimation method for resource selection. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 298–305, New York, NY, USA, 2003a.

Si, L. and Callan, J. A semisupervised learning method to merge search engine results. *ACM Transactions on Information Systems*, 21(4):457–491, 2003b.

Si, L., Jin, R., Callan, J., and Ogilvie, P. A language modeling framework for resource selection and results merging. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, pages 391–397, New York, NY, USA, 2002.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Association for Computational Linguistics, Stroudsburg, PA, USA, 2008.

Sohn, T., Li, K. A., Griswold, W. G., and Hollan, J. D. A diary study of mobile information needs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 433–442, New York, NY, USA, 2008.

Sushmita, S., Joho, H., and Lalmas, M. A task-based evaluation of an aggregated search interface. In *Proceedings of the 16th International Symposium on String Processing and Information Retrieval*, SPIRE '09, pages 322–333, Springer-Verlag, Berlin, Heidelberg, 2009.

Sushmita, S., Joho, H., Lalmas, M., and Jose, J. M. Understanding domain relevance in web search. In *WWW Workshop on Web Search Result Summarization and Presentation*, 2010a.

Sushmita, S., Joho, H., Lalmas, M., and Villa, R. Factors affecting click-through behavior in aggregated search interfaces. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 519–528, New York, NY, USA, 2010b.

Sushmita, S., Piwowarski, B., and Lalmas, M. Dynamics of genre and domain intents. In *Proceedings of the 6th Asia Information Retrieval Societies Conference*, AAIRS '10, pages 399–409, Springer-Verlag, Berlin, Heidelberg, 2010c.

Sutton, R. S. and Barto, A. G. *Introduction to Reinforcement Learning.* MIT Press, Cambridge, MA, USA, 1st edition, 1998.

Thomas, P. and Shokouhi, M. Sushi: Scoring scaled samples for server selection. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 419–426, New York, NY, USA, 2009.

Treeratpituk, P. and Callan, J. Automatically labeling hierarchical clusters. In *Proceedings of the 2006 International Conference on Digital Government Research*, DG.O '06, pages 167–176, Digital Government Society of North America. 2006.

Trippas, J. R. Spoken conversational search: Information retrieval over a speech-only communication channel. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 1067–1067, New York, NY, USA, 2015.

Tsur, G., Pinter, Y., Szpektor, I., and Carmel, D. Identifying web queries with question intent. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 783–793, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2016.

Turpin, L., Kelly, D., and Arguello, J. To blend or not to blend? perceptual speed, visual memory and aggregated search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, New York, NY, USA, 2016.

van Rijsbergen, C. J. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

Wang, C., Liu, Y., Zhang, M., Ma, S., Zheng, M., Qian, J., and Zhang, K. Incorporating vertical results into search click models. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 503–512, New York, NY, USA, 2013.

Wang, Y., Yin, D., Jie, L., Wang, P., Yamada, M., Chang, Y., and Mei, Q. Beyond ranking: Optimizing whole-page presentation. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 103–112, New York, NY, USA, 2016.

Wen, J.-R., Nie, J.-Y., and Zhang, H.-J. Clustering user queries of a search engine. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 162–168, New York, NY, USA, 2001.

Wu, W.-C., Kelly, D., Edwards, A., and Arguello, J. Grannies, tanning beds, tattoos and nascar: Evaluation of search tasks with varying levels of cognitive complexity. In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX '12, pages 254–257, New York, NY, USA, 2012.

Xu, J. and Li, H. Adarank: A boosting algorithm for information retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 391–398, New York, NY, USA, 2007.

Xue, X.-B., Zhou, Z.-H., and Zhang, Z. M. Improving web search using image snippets. *ACM Transactions of Internet Technology*, 8(4):21:1–21:28, 2008.

Yamamoto, T., Liu, Y., Zhang, M., Dou, Z., Zhou, K., Markov, I., Kato, M. P., Ohshima, H., and Fujita, S. Overview of the ntcir-12 imine-2 task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, NTCIR '16, National Institute of Informatics. 2016.

Yue, Y., Patel, R., and Roehrig, H. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1011–1018, New York, NY, USA, 2010.

Zhou, K., Cummins, R., Halvey, M., Lalmas, M., and Jose, J. M. Assessing and predicting vertical intent for web queries. In *Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR'12, pages 499–502, Springer-Verlag, Berlin, Heidelberg, 2012a.

Zhou, K., Cummins, R., Lalmas, M., and Jose, J. M. Evaluating reward and risk for vertical selection. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2631–2634, New York, NY, USA, 2012b.

Zhou, K., Cummins, R., Lalmas, M., and Jose, J. M. Evaluating aggregated search pages. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 115–124, New York, NY, USA, 2012c.

Zhou, K., Cummins, R., Lalmas, M., and Jose, J. M. Which vertical search engines are relevant? In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 1557–1568, New York, NY, USA, 2013a.

Zhou, K., Lalmas, M., Sakai, T., Cummins, R., and Jose, J. M. On the reliability and intuitiveness of aggregated search metrics. In *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '13, pages 689–698, New York, NY, USA, 2013b.

Zhou, K., Demeester, T., Nguyen, D., Hiemstra, D., and Trieschnigg, D. Aligning vertical collection relevance with user intent. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1915–1918, New York, NY, USA, 2014.

Zhuang, J., Mei, T., Hoi, S. C., Xu, Y.-Q., and Li, S. When recommendation meets mobile: Contextual and personalized recommendation on the go. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 153–162, New York, NY, USA, 2011.