ON THE EVALUATION OF AGGREGATED WEB SEARCH

KE ZHOU

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
Doctor of Philosophy

SCHOOL OF COMPUTING SCIENCE
COLLEGE OF SCIENCE AND ENGINEERING
UNIVERSITY OF GLASGOW

JULY 2014

Abstract

Aggregating search results from a variety of heterogeneous sources or so-called *verticals* such as news, image and video into a single interface is a popular paradigm in web search. This search paradigm is commonly referred to as *aggregated search*. The heterogeneity of the information, the richer user interaction, and the more complex presentation strategy, make the evaluation of the aggregated search paradigm quite challenging. The *Cranfield paradigm*, use of test collections and evaluation measures to assess the effectiveness of information retrieval (IR) systems, is the de-facto standard evaluation strategy in the IR research community and it has its origins in work dating to the early 1960s. This thesis focuses on applying this evaluation paradigm to the context of aggregated web search, contributing to the long-term goal of a complete, reproducible and reliable evaluation methodology for aggregated search in the research community.

The Cranfield paradigm for aggregated search consists of building a test collection and developing a set of evaluation metrics. In the context of aggregated search, a test collection should contain results from a set of verticals, some information needs relating to this task and a set of relevance assessments. The metrics proposed should utilize the information in the test collection in order to measure the performance of any aggregated search pages. The more complex user behavior of aggregated search should be reflected in the test collection through assessments and modeled in the metrics.

Therefore, firstly, we aim to better understand the factors involved in determining relevance for aggregated search and subsequently build a reliable and reusable test collection for this task. By conducting several user studies to assess vertical relevance and creating a test collection by reusing existing test collections, we create a testbed with both the vertical-level (user orientation) and document-level relevance assessments. In addition, we analyze the relationship between both types of assessments and find that they are correlated in terms of measuring the system performance for the user.

Secondly, by utilizing the created test collection, we aim to investigate how to model the aggregated search user in a principled way in order to propose reliable, intuitive and trustworthy evaluation metrics to measure the user experience. We start our investigations by studying solely evaluating one key component of aggregated search: vertical selection, i.e. selecting the relevant verticals. Then we propose a general utility-effort framework to evaluate the ultimate aggregated search pages. We demonstrate the fidelity (predictive power) of the proposed metrics by correlating them to the user preferences of aggregated search pages. Furthermore, we meta-evaluate the reliability and intuitiveness of a variety of metrics and show that our proposed aggregated search metrics are the most reliable and intuitive metrics, compared to adapted diversity-based and traditional IR metrics.

To summarize, in this thesis, we mainly demonstrate the feasibility to apply the Cranfield Paradigm for aggregated search for reproducible, cheap, reliable and trustworthy evaluation.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my two co-advisors Dr. Mounia Lalmas and Prof. Joemon M. Jose for the continuous support of my Ph.D study and research, for their patience, motivation, enthusiasm, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study. In addition, I would give a big thank to Dr. Ronan Cummins, not only for the guidance at the early stage of my PhD study, but also all the Friday drinks and fun that are embedded in my memory.

I would like to thank Dr. Tetsuya Sakai for giving me the chance for collaborating with Microsoft Research and thank Microsoft colleagues for a warm hospitality in Beijing. Thanks for offering me the summer internship opportunities in their groups and leading me working on diverse exciting projects. A special thank to Dr. Tetsuya Sakai and Dr. Zhicheng Dou for their mentoring and valuable research discussions during my internship. Many thanks to Dr. Haitao Yu for kindly collaborating with me and offering helps during the internship.

I could not have survived living here without a great support from the Glasgow and ex-Glasgow teams. I thank my good friends Dr. Martin Halvey, Stefan Raue and Dr. Robert Villa for all their support during my PhD. I learned a lot from you guys. I also thank my fellow labmates in Glasgow Information Retrieval Group: Stewart Whiting, Jesus Rodriguez Perez, James McMinn, Philip MacParlane, Horatiu Bota for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last four years.

Before joining Glasgow IR group, I tried my first steps in doing research with my mentor at Northeastern University, Prof. Jingbo Zhu. I thank Prof. Jingbo Zhu, in particular for encouraging me to pursue a PhD.

Finally, I would like to thank my family for their love and moral support, and especially my parents and Tian for inspiration and always being there when I needed.

Dedication to My Parents & Tian!

Contents

List of Tables

List of Figures

Part I

Introduction and Background

1 Introduction

Information Retrieval (IR) systems are common place and are vital in many facets of modern life for users to search information to fulfill their information needs. For example, on the World Wide Web, it has become prevalent that users treat search engines as their starting point for retrieving or browsing the web. It has been reported [ses] in 2012 that the leading commercial web search engine indexes over 30 trillion unique web pages (URLs) and answers a total of one hundred million search queries each month. It is not only the scale of the search engine that is increasing, users' information needs and the search engines themselves are becoming increasingly diverse as well.

Due to the increasing heterogeneity of the web environment, diverse contents are available on the web while their contents differ in terms of media (e.g. text, image, video, audio), genre (e.g. encyclopedia/wiki, FAQs, news, blogs) and topicality (entertainment, sports, technology). To access all those disparate sources of information, there exists a variety of so-called *search verticals* operated by the commercial search company and each vertical search engine specializes in searching one specific type of information, for example, news search, image search and video search. Note that the information retrieved across all those verticals could have significantly different characteristics, for example, they could be across different media, genres or topicality.

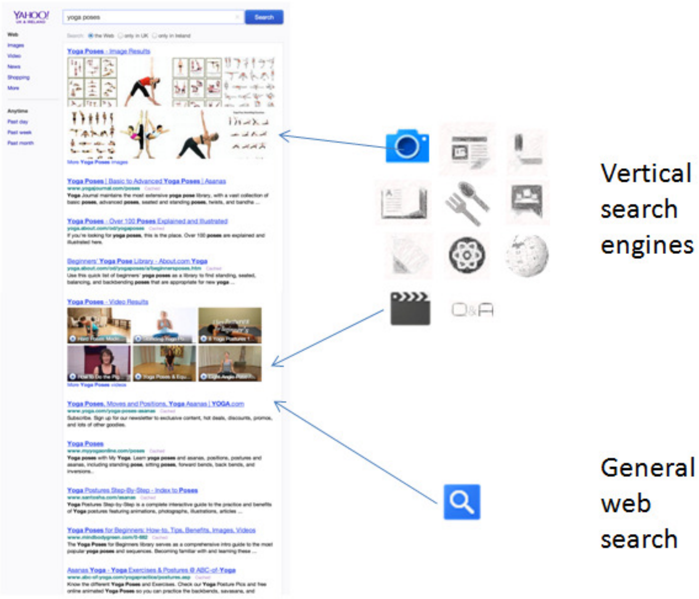
Given the huge amount of heterogeneous information available, users' information needs are becoming increasingly more diverse as well. Users tend to search across all different aspects relating to their lives, for example, learning how to perform yoga poses, checking the weather of the city, solving a task on how to copy and paste using the word processor, etc. Those information needs and their corresponding issued user search queries generally could refer to different search verticals. Actually, recent research [Arguello et al., 2009] has shown that 74% of the queries were assigned at least one relevant vertical(s)[1] (in addition to "General Web" search engine) by the human editors. Until recently each type of content was dealt with in a separate way through search verticals, and users switched between verticals to access information of a given type.

Despite of the usefulness of each individual vertical search engine, however, in a survey [Jup, 2008] carried out in 2008, it indicates that 35% of users do not use vertical search. As we have demonstrated above, this does not mean that users do not issue queries with one or more vertical intents. Rather, it means that users do not often switch from their default search engine (i.e. "General Web" search engine) to other vertical

[1] In this section, we regard the query has that vertical intent if we specify the vertical is relevant to the query.

search engines. In other words, it implies that they expect the most relevant results to be returned from their default search engine to satisfy their information needs.

To remedy the fact that vertical search is not prevalently used by the users, a new search paradigm, *aggregated search*, has been proposed in both the research community [Murdock and Lalmas, 2008] and the commercial world [Arguello et al., 2009]. Basically, aggregated search deals with aggregating search results from those different search verticals and present them alongside with organic "General Web" search results into one single interface. Currently, this aggregated search paradigm[Murdock and Lalmas, 2008] has been implemented by most major commercial search engines. An example of one commercial aggregated search system is shown in Figure 1.1. From the figure, we can observe that for the user that aims to learn to practice yoga (issued with the query "yoga poses"), not only "General Web" results, but "Video" and "Image" vertical results are also aggregated for the final presentation on the search engine result page (SERP).



yoga poses

Yoga Poses - Image Results

More Yoga Poses images

Yoga Poses | Basic to Advanced Yoga Poses | Asanas
www.yogajournal.com/poses Cached
Yoga Journal maintains the most extensive yoga pose library, with a vast collection of basic poses, advanced poses, seated and standing poses, twists, and bandha ...

Yoga Poses - Over 100 Poses Explained and Illustrated
yoga.about.com/od/yogaposes Cached
If you're looking for yoga poses, this is the place. Over 100 poses are explained and illustrated here.

Beginners' Yoga Pose Library - About.com Yoga
yoga.about.com/od/yogaposes/a/beginnersposes.htm Cached
Use this quick list of beginners' yoga poses as a library to find standing, seated, balancing, and backbending poses that are appropriate for new yoga ...

Yoga Poses - Video Results

More Yoga Poses videos

Yoga Poses, Moves and Positions, Yoga Asanas | YOGA.com
www.yoga.com/yoga-poses-asanas Cached
Subscribe. Sign up for our newsletter to exclusive content, hot deals, discounts, promos, and lots of other goodies.

Yoga Poses
www.myyogaonline.com/poses Cached
Yoga poses with My Yoga. Learn yoga poses and asanas, positions, postures and asanas, including standing pose, sitting poses, forward bends, back bends, and inversions.

Yoga Postures Step-By-Step - Index to Poses
www.santosha.com/asanas Cached
Yoga Postures Step-by-Step is a complete interactive guide to the practice and benefits of Yoga postures featuring animations, photographs, illustrations, articles ...

Yoga Poses for Beginners: How-to, Tips, Benefits, Images, Videos
www.mindbodygreen.com/0-682 Cached
The Yoga Poses for Beginners library serves as a comprehensive intro guide to the most popular yoga poses and sequences. Becoming familiar with and learning these ...

Asanas Yoga - Yoga Exercises & Postures @ ABC-of-Yoga
www.abc-of-yoga.com/yogapractice/postures.asp Cached
Know the different Yoga Poses and Exercises. Check our Yoga Posture Pics and free online animated Yoga Poses so you can practice the backbends, savasana, and

Figure 1.1: Example of an commercial aggregated search result page addressing the user information need that with issued query "yoga poses". In accompanying with "General Web" organic search results, results from two search verticals, "Video" and "Image", are also selected and returned.

Aggregated search is related to the *federated search* [Shokouhi and Si, 2011] problem in the IR research community. Federated search is the problem of automatically searching across multiple distributed collections or resources. Federated search can and has been applied in different environments, such as meta-search [Meng et al., 2001], and personal search [Kim and Croft, 2010]. In particular, aggregated search can be treated as one special case of the federated search problem where the focus is on federating *heterogeneous* search verticals available on the web. Aggregated search deals with such diverse and heterogeneous information spaces and interaction data. It cannot be assumed that previous federated research holds in such a diverse scenario.

There are three main differences that distinguish aggregated search from traditional federated search task: (1). heterogeneity; (2). richer interaction and (3). more complex presentation strategies. Different from much of the research into traditional federated search [Shokouhi and Si, 2011] that has utilized documents of the same type (i.e. homogeneous), and do not investigate a truly rich and diverse information space, aggregated search focuses on the web context, with various existing vertical search engines (Blog, News, Image) available. Those verticals differ significantly among each other (with respect to the item type and their ranking functions). In aggregated search, users' interaction data (such as query logs) can provide a rich source of information for learning user behavior which is normally not applicable in the traditional federated search domain. In addition, presenting heterogeneous information is more complex than the typical single ranked list (e.g. ten blue links) employed in homogeneous ranking in traditional federated search. There are three main types of presentation designs for aggregated search: (1) results from the different verticals are blended into a single list (of blocks), referred to as *blended*; (2) results from each vertical are presented in a separate (e.g. horizontal paralleled) panel (tile), referred to as *non-linear blended*; and (3) vertical results can be accessed in separate tabs, referred to as *tabbed*. A combination of all three is also possible.

Aggregation can be useful in other context as well, such as personal search. Recent research [Kim and Croft, 2010] has shown that on the personal machines, users maintain information needs to search for different types of information (e.g. email, calendar, articles, codes, etc.). A set of research prototypes [Kim and Croft, 2010] and commercial softwares (e.g. Spotlight) are developed for this information federation. Results in Spotlight are presented in a ranked lists, with each type presented in a separate block. The objective of selection and ranking of each block is to satisfy the users with least effort to find what he/she intended.

In the remainder of this thesis, when we refer to aggregated search, we focus on the area of aggregated web search (i.e. aggregating search verticals available on the web), that has not been deeply investigated in prior work until recently [Arguello and Fernando, 2013]. The *verticals* we consider in this thesis are the most prevalently used vertical search engines developed by the commercial web search engine companies (such as Google, Yahoo and Bing), varying across different media, genres and topicality. Examples of such vertical search engines include multimedia vertical search engines (Image and Video), topicality-specific vertical search engines (Recipe, Shopping and Scholar) and genre-specific search engines (News, Books, Blog, Answer, Wiki and Discussion). In this thesis, we also mainly investigate aggregated search in the desktop environment (e.g. presentation strategies and user models more specifically tailored towards desktop search users). Aggregated search in the mobile environment, with devices of limited screen sizes and other user contexts can be exploited (such as geo) rather than the search query, is out of the scope of this thesis.

As the management thinker Peter Drucker says,

> *"If you can't measure it, you can't improve it."*

evaluation is an important concerns in the IR community. Although numerous work exists for IR evaluation, little work has been conducted to comprehensively evaluate in the context of aggregated search (except recent endeavors [Arguello et al., 2011b]). The main aim of this thesis is to tackle the evaluation of aggregated search, i.e. given any two aggregated search pages for an information need, we would like to tell which one is preferred by the user. Compared with IR evaluation, aggregated search evaluation is more challenging due to the new problems it poses that we have mentioned. Firstly, it is challenging due to the introduction of another dimension for evaluation (types of information, i.e. verticals where the information originated from). User's orientation to the verticals for a given information need has been shown to affect user's interaction patterns [Sushmita et al., 2010]. In addition, since there could be results coming from multimedia other than text (e.g. image, video), this could change the user's interaction pattern with the presented search page. Thirdly, because of the different presentation strategies involved, different interaction patterns could be expected. All the above challenges motivate this thesis.

Current IR evaluation work fall into one of two categories: system-based evaluation and user-centered evaluation approaches. System-based evaluation can be also referred to as the Cranfield paradigm [Sanderson, 2010] where a test collection is constructed in order to conduct the evaluation. A test collection typically consists of a set of documents, a set of information needs (topics) and a set of relevance judgments between a given document and a topic, collected from one or a few assessors. Following that, for a given ranked list, the relevance label of the document-topic pairs are utilized to infer the utility of the ranked list. The benefits of applying system-based evaluation lies in the fact that the experiments can be easily reproduced and interpreted. However, the arguments against this Cranfield paradigm is that it lacks a model of real-world "user behavior". Note that the system-based evaluation is normally conducted at the early stage of the development and evaluation cycle when the system has not deployed been to be used by a large population of real users. Intensive evaluations of various system algorithms can be iteratively performed in order to achieve the optimal performance for system deployment. However, the users' search success and satisfaction with a search system may not be always accurately reflected by the system-based evaluation metrics [Turpin and Scholer, 2006]. In addition, the assessments provided within the Cranfield paradigm are collected in an offline fashion with a set of assessors (potentially a small number of expert users). Those assessors provide their explicit assessments (e.g. assessing the relevance between the information need and the document) or even their search preference (e.g. personal preference towards specific information disregarding the information need). Those assessments are made to be sufficiently general so that they can be used to evaluate the various simulated situations that are supposed to be correlated with real user satisfaction.

On the other hand, user-centered evaluation, such as A/B test [Kohavi et al., 2009] and inter-leaving [Radlinski et al., 2008] approaches, aims to measure the performance of the system by deploying real-systems that are interacted by the real users. By inferring from the user implicit interaction data (such as query and click logs), one can make conclusions or implications of the users' preference of the deployed systems. Different from the system-based evaluation, the pros of this user-centered approach is that it involves with real users and real search tasks. In addition, if collecting explicit user feedback, more information can be obtained on "why" a given system is better than the other. Despite those benefits of conducting user experiments, the user-centered evaluation methodology suffers from the fact that it can not be reproduced easily (although simulated evaluation could be conducted [Ponnuswami et al., 2011a]) and it has to be more user-aware (i.e. the researchers have to be more careful when applying potentially harmful changes to the IR system, in case the real user experience of the system could be degraded).

1.1 Thesis Statement

The broad question that motivates the research in this thesis is: *Can we adapt the traditional Cranfield paradigm evaluation approach (i.e. test collection based evaluation) to measure the performance of the aggregated search systems?* Although the Cranfield paradigm has been extensively investigated [Sanderson, 2010] [2] and there exists some work in evaluating aggregated search either in the offline [Arguello et al., 2009] or online fashion [Ponnuswami et al., 2011b], however, evaluating aggregated search using a TREC style evaluation fashion has yet been investigated. This thesis aims to close some of these gaps, contributing to the long-term goal of a complete, reproducible evaluation approach for aggregated search in the research community.

The statement of this thesis is that the complexity of aggregated search user behavior can be modeled for evaluation and the adapted Cranfield paradigm could be a reliable and reusable approach to measure aggregated search system performance. The main aim of this thesis is to apply the Cranfield paradigm to aggregated search. This consists of building a test collection and developing a set of metrics. A test collection should contain results from a set of verticals, some information needs relating to this task and relevance assessments. The metrics proposed utilized the information in the test collection in order to measure the performance of any aggregated search pages. The user behavior of aggregated search should be reflected and modeled in the metrics.

In particular, by studying a variety of approaches to assess vertical relevance, we demonstrate that reliable vertical-level relevance (i.e. users' type preference) can be obtained with relatively low cost. Moreover, by reusing existing web-based collections, a reliable aggregated search test collection can be built very cheaply, requiring only a small amount of document level relevance assessments. Furthermore, by utilizing the vertical and document relevance assessments available in the test collection, we show that users' behavior on the (blended) aggregated search page can be modeled and be utilized in the evaluation metrics to measure aggregated search system performance (either in part or the whole system). Finally, we present that our proposed evaluation metrics are *reliable*, *intuitive* and *trustworthiness*. The *reliability* refers to the ability of the metrics to statistically discriminate aggregated search systems while the *intuitiveness* focuses on whether the metrics perform sufficiently intuitive to capture different key components of aggregated search? The *trustworthiness* refers to the relatively strong *predictive power* of the metrics [Sanderson et al., 2010] to correlate with human's preference of aggregated search pages.

[2] Actually, such evaluation methodology has been widely used in a variety of search tasks in evaluation forums such as TREC [tre, a], NTCIR [ntc] and CLEF [cle], due to its reproducibility and ease of comparison among groups of researchers around the world.

At the time of publication of this thesis, TREC has launched a new search task FedWeb ("Federated Search task" [Demeester et al., 2013]) in 2013 and 2014. Part of this thesis (e.g. Chapter 5) has also contributed to some of the choices made in the evaluation setting in TREC FedWeb 2014 task.

1.2 Research Outline and Questions

An overview of all the research conducted in this thesis is shown in Figure 1.2. The work and research questions can be split mainly into two parts: assessments (test collection) and metrics (evaluation measures). The main objective of this thesis is to model the most essential components of the complex user behavior on aggregated search pages and incorporate them into both the assessments and the evaluation metric space. Firstly, we aim to understand better on how to make *assessments* for aggregated search and therefore create a sufficient, reliable and reusable test collection for this task. Secondly, we aim to investigate how to model the aggregated search user in a principled way to propose reliable, intuitive and trustworthy evaluation metrics to measure the user experience.

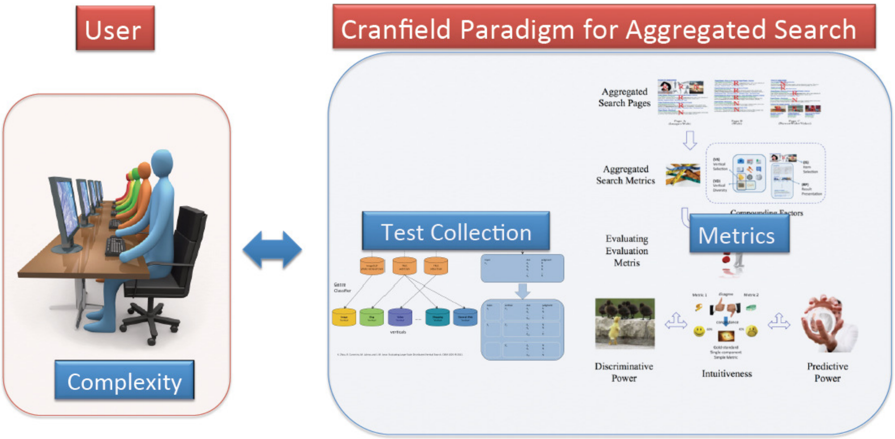


Figure 1.2: Overview of the thesis.

1.2.1 Aggregated Search Assessments

As *vertical* is one crucial component that has been introduced in aggregated search (compared to other traditional IR tasks), we start our investigations by studying whether different underlying assumptions made for vertical relevance assessments affect a user's perception of the relevance of verticals. Although some prior work [Arguello et al., 2009, Ponnuswami et al., 2011b] exist on assess vertical relevance, either by explicitly asking annotators or implicit deriving from use behavior data, it is unclear on the impact of different assumptions made to the final assessments. Therefore, we formalize the different assumptions made by prior work, design and conduct several user studies in order to answer the following questions:

> **RQ 1**: Are there any differences between the assessments made by users from a pre-retrieval user-need perspective (viewing only vertical labels prior to seeing the final SERP) and the assessments made by users from a post-retrieval user perspective (viewing the vertical results in the final SERP)?
>
> **RQ 2**: When using "general web" results as a reference for making vertical relevance assessments, are these assessments able to predict the users' pairwise preference between any two verticals?
>
> **RQ 3**: Does the context (results returned from other verticals) in which vertical results are presented affect a user's perception of the relevance of the vertical of interest?
>
> **RQ 4**: Is the vertical preference information provided by a population of users able to predict the "perfect" embedding position of a vertical?

In answering these questions, we identify that previous vertical assessment assumptions vary on three types of variants in the assessments: influencing factors, vertical relevance dependency and the assessment grades. We conduct user studies to investigate on each of these.

After analyzing how to gather vertical-level relevance assessments, we aim to build a TREC-style aggregated search test collection that collects document-level relevance assessments. This TREC-style aggregated search test collection should consist of a set of verticals, each populated by a set of items (documents), a set of topics (information needs) related to one or multiple verticals, and a set of relevance assessments between any pair of topic and item. Since it requires huge manual effort to collect document-level relevance assessments, we aim to build a test collection by re-using a current web-based test collection that has been used in TREC web track [tre, b]. We aim to address the following research questions:

> **RQ 5**: Can we reuse existing test collections to construct a test collection for aggregated search?
>
> **RQ 6**: Is the constructed test collection reliable? What is the impact of misclassification (of items into verticals) to the evaluation of systems?

After constructing an aggregated search test collection with both vertical-level and document-level relevance assessments, we turn to study whether there is some correlation between vertical-level assessments (from the pre-retrieval user intent perspective) and the one obtained by analyzing the distribution of relevance within the vertical. We design a set of approaches to derive vertical relevance from document relevance and conduct several experiments to investigate:

> **RQ 7**: Can the vertical relevance be derived from document relevance judgments and therefore ranked similarly to the user vertical intent (orientation)?

> **RQ 8**: Can we appropriately threshold the derived vertical rankings and ultimately align them with the binary vertical selection decision made by the users?

We found that collection-based vertical relevance derived from document-level relevance assessments can be aligned relatively well with users' vertical intent. Therefore this suggests that collection-based vertical relevance can be utilized as an approximate surrogate for measuring user's vertical intent (orientation), and vice versa.

1.2.2 Aggregated Search Evaluation Metrics

After addressing the above questions regarding the assessments for aggregated search that should be used to sufficiently model the user to build a test collection, we turn to the other key component for Cranfield paradigm evaluation: the evaluation metrics. Given the vertical-level and document-level assessments, and different stages of aggregated search, we investigate how to effectively measure the system performance.

Firstly, we study the evaluation of another key component of aggregated search: vertical selection. Generally, it is not always the case that providing additional results from other verticals can benefit the users. Only selecting relevant verticals that a large population of users that are favoring can be rewarding while selecting irrelevant verticals that not a lot of users intended can result in the risks of hurting the user experience. Therefore, we propose the risk-aware vertical selection metric and we utilize it to study a number of vertical selection approaches:

> **RQ 9**: For evaluating vertical selection, rather than solely consider reward (selecting relevant verticals), can we measure the performance on maximising reward while minimising risk (selecting irrelevant verticals)?

> **RQ 10**: How effective and robust are existing vertical selection approaches considering the varying types of user (risk-averse and risk-seeking)?

Following the evaluation of vertical selection, we turn to a more thorough evaluation of the entire aggregated search system, i.e. measuring the effectiveness of the ultimate aggregated search pages. We formalize the layout of the blended aggregated search page and propose a utility-effort evaluation framework to capture the user behavior in order to answer the following questions:

> **RQ 11**: Do users agree with each other when assessing the preference of aggregated search pairs?

> **RQ 12**: Can we evaluate aggregated search pages (the whole aggregated search system) in a way that capture both effort and utility (relevance) formally? How can we utilize (combine) both vertical relevance and document relevance when evaluating aggregated search pages?

> **RQ 13**: Do those aggregated search metrics possess strong *predictive power*, i.e. aligning with the real user preference of aggregated search pages?

> **RQ 14**: Can we personalize the evaluation based on each types of user?

In addressing the above questions, we propose a set of aggregated search metrics and demonstrate that they have relatively strong predictive power to align with the users' preference of aggregated search pages.

Finally, as there are multiple key components in aggregated search, namely, vertical selection, vertical diversity, item selection and result presentation, for all different types of metrics, we aim to understand how well different metrics perform regarding both their reliability and intuitiveness:

> **RQ 15**: How do all different suites of metrics (traditional IR, diversity IR and aggregated search) perform with respect to reliability, i.e. the ability to statistically discriminate aggregated search systems?
>
> **RQ 16**: Are all different suites of metrics perform sufficiently intuitive to capture different key components of aggregated search?

1.3 Summary of Contributions

The work on aggregated search evaluation to information retrieval is a relatively new field of research. The main contributions lies in our practical contribution of the created test collection and collected assessments, and our theoretical contribution on the evaluation metric frameworks.

C1. We are the first work that comprehensively understand and compare different vertical relevance assessment processes and assumptions. In particular, we employ all of the various strategies present in the literature and design extensive user studies to collect and study the correlation of those vertical relevance assessments.

C2. We are the first work to propose methodology of reuse to construct a test collection for aggregated search from existing test collections. We build a practical, reliable and large-scale test collection for aggregated search by using a SVM classifier to classify items into various verticals (types). This created test collection can be beneficial for the IR research community.

C3. We extensively study different approaches to derive collection-based vertical relevance in the context of over a hundred heterogeneous search engines. The scale and diversity of the search engines used has not been studied previously. We also conduct a user study to verify that the collection-based vertical relevance derived from document judgments can be aligned with the user vertical intent.

C4. We propose a new risk-aware VS evaluation metric. Rather than treating a vertical as either relevant or irrelevant given a query, as mostly done in current work, we propose a general framework to evaluate the reward and risk for vertical selection on a per user basis. We also perform an analysis of the effectiveness and robustness of several vertical selection approaches across different types of user with varying level of risk-tolerance.

C5. We propose a general framework for the evaluation of aggregated search pages that capture both effort and reward in a formal way. We also outline a novel approach for simulating aggregated search pages and collect a large set of user preferences over page pairs. With the user preference data, we show that our proposed metrics can be aligned with the user and can be further improved by personalising for each type of user.

C6. Our work is the first endeavour to study the reliability and intuitiveness of aggregated search metrics extensively. We also propose a framework to quantify the intuitiveness of each evaluation metric to capture the key component(s) of aggregated search.

1.4 Origins of the Materials

Most of the material presented in this thesis has previously appeared in several conference papers published in the course of this PhD programme:

P1. K. Zhou, R. Cummins, M. Lalmas and J. M. Jose. Evaluating Large-Scale Distributed Vertical Search, CIKM Workshop on Large-Scale and Distributed Information Retrieval (LSDS-IR 2011), Glasgow, Scotland, UK, 2011.
[This publication is included in Chapter 4.]

P2. K. Zhou, R. Cummins, M. Halvey, M. Lalmas and J. M. Jose. Assessing and Predicting Vertical Intent for Web Queries, 34th European Conference on Information Retrieval (ECIR 2012), Barcelona, Spain, 2012.
[This publication is included in Chapter 3.]

P3. K. Zhou, R. Cummins, M. Lalmas and J. M. Jose. Evaluating Aggregated Search Pages, 35th ACM Conference on Special Interest Group on Research and Development in Information Retrieval (SIGIR 2012), Portland, USA, 2012.
[This publication is included in Chapter 7.]

P4. K. Zhou, R. Cummins, M. Lalmas and J. M. Jose. Evaluating Reward and Risk for Vertical Selection, 21st ACM International Conference on Information and Knowledge Management (CIKM 2012), Hawaii, USA, 2012.
[This publication is included in Chapter 6.]

P5. K. Zhou, R. Cummins, M. Lalmas and J. M. Jose. Which Vertical Search Engines are Relevant? Understanding Vertical Relevance Assessments for Web Queries, International World-Wide Web Conference (WWW 2013), Rio de Janeiro, Brazil, 2013.
[This publication is included in Chapter 3.]

P6. K. Zhou, T. Sakai, M. Lalmas, Z. Dou and J. M. Jose. Evaluating Heterogeneous Information Access (Position Paper), SIGIR Workshop on Modeling User Behavior for Information Retrieval Evaluation (MUBE 2013), Dublin, Ireland, 2013.
[This publication is included in Chapter 9.]

P7. K. Zhou, M. Lalmas, T. Sakai, R. Cummins and J. M. Jose. On the Reliability and Intuitiveness of Aggregated Search Metrics, ACM International Conference

on Information and Knowledge Management (CIKM 2013), San Francisco, USA, 2013.
[This publication is included in Chapter 8.]

P8. K. Zhou, T. Demeester, D. Nguyen, D. Hiemstra and D. Trieschnigg. Aligning Vertical Collection Relevance with User Intent. ACM International Conference on Information and Knowledge Management (CIKM 2014), Shanghai, China, 2014.
[This publication is included in Chapter 5.]

1.5 Organisation

The thesis is divided into four main parts. Part I presents motivations, research questions, technical background and the state-of-the-art. Part II-III presents our proposed Cranfield paradigm evaluation methodology for aggregated search, including the assessments (Part II) and the evaluation metrics (Part III). Part IV concludes and outlooks the future work. The detailed organization of the thesis is outlined below.

Part I. Introduction and Background

Chapter 1 includes this introduction that mainly describes the motivation of this work. Research questions and methodology are also discussed in this chapter.

Chapter 2 presents prior work in this area. This consists of a brief overview of current IR evaluation methodology, followed by the descriptions of different state-of-the-art aggregated and federated search approaches.

Part II. On the Assessments of Aggregated Search

Chapter 3 presents several designed user study to understand vertical relevance assessments.

Chapter 4 describes and evaluates our reuse approach to create an aggregated search test collection.

Chapter 5 presents and evaluates various approaches to derive vertical relevance from document level relevance assessments.

Part III. On the Evaluation Metrics of Aggregated Search

Chapter 6 describes our risk-aware vertical selection metric and utilizes it to evaluate a variety of vertical selection approaches.

Chapter 7 presents our proposed utility-effort evaluation framework for aggregated search. We also evaluate the proposed metrics based on their predictive power.

Chapter 8 discusses the reliability and intuitiveness of a variety of aggregated search metrics to capture aggregated search key component(s).

Part IV. Conclusions and Discussions

Chapter 9 concludes the thesis by discussing the main findings, implications for each research question and presents the future work that could be carried out.

Readers familiar with IR evaluation and federated/aggregated search approaches can skip over Chapter 2.

2
Prior Work

In this chapter, we introduce the background and the most relevant prior work to this thesis. Therefore, we aim to cover the two main themes of this thesis (as indicated by the thesis title): (IR) evaluation (Section 2.1) and aggregated search (Section 2.2). With respect to IR evaluation, most immediately relevant literatures are the test collection evaluation methodology widely used in the IR community [Sanderson, 2010] and we cover the most essential materials in Section 2.1.1. We also briefly discuss the other alternatives in Section 2.1.2, i.e. the lab-based user studies [Ruthven and Kelly, 2013], the online evaluation approaches (A/B testing [Kohavi et al., 2009] and inter-leaving [Radlinski et al., 2008]). With respect to aggregated search, we review approaches for each of the key component of aggregated search, followed by a discussion of evaluation settings carried out within the prior work. Within the entire literature review, we mainly focus on the approaches that are most relevant while briefly discussing the broad background. We refer to the readers other relevant books or surveys if more details are needed.

2.1 IR Evaluation

Evaluation is central in development of IR community. The performance of search systems can be evaluated with *user-oriented* and *system-oriented* measures. The former are obtained through user studies conducted to examine and reflect upon various aspects of user behaviours. The latter relies on reusable test collections (i.e. document, query and relevance judgements) to assess the search quality. We review previous work in details on those two types of endeavors respectively in the following sections.

2.1.1 System-oriented Evaluation

System-oriented evaluation can be also referred to as Cranfield paradigm in the IR community. Basically, rather than involving users to directly make judgements on the system performance, the evaluation is conducted by using a test collection. The benefits of doing so is that the experiments can be reusable and reproducible.

Test Collection

Test collection has the following components: a static collection of retrievable documents, a set of topics (user information needs) with topic descriptions that define what should and should not be considered relevant, a set of relevance assessments for all topic-document pairs. Usually, the most time-consuming part for building the test collection is to collect the relevance judgments, especially for the large-scale collection with millions or even billions of documents (e.g. ClueWeb [clu, 2009]). Therefore, it is not possible to judge all the documents for a given topic.

Due to the fact that most documents are not relevant, instead, assessors typically judge only those documents most likely to be relevant. These can be determined using a method that is called pooling [Jones and van Rijsbergen, 1975] in the research community. The basic idea is to take the union of top results from a wide range of systems. Documents within this set are judged and documents outside of this set are automatically assumed to be non-relevant. It has been demonstrated that pooling is a reliable methodology for building a test collection [Zobel, 1998] and has been widely adopted in a variety of evaluation forums (e.g. TREC).

Traditional Metrics

Based on the relevance assessments within the test collection, a suite of evaluation metrics operate on a ranking of known relevant/non-relevant documents to determine the quality of the ranking and assume that correlates to the ultimate user experience of the search system [Sanderson et al., 2010].

Traditional IR evaluation is based on topical relevance, $qrel(q, d)$, between a query q and a document d. Traditional IR metrics ignore the document type (e.g. vertical) and measure the quality of a ranked list l by modelling the gain $G@l$ of a user reading all documents in that list l. Perhaps the most basic metrics associated with retrieval effectiveness are precision and recall [Baeza-Yates et al., 1999]. $P@k$ (precision at rank k) assumes that after reading the top k results in l, a user's gain $G@k$ solely depends on the number of relevant documents within the top k results. $R@k$ (recall at rank k) instead assumes that after reading the top k results in l, a user's gain $G@k$ solely depends on the fraction of relevant retrieved documents over all the relevant documents in the whole collection (C_R). As observed by Cleverdon [Sanderson, 2010], precision and recall often have an inverse relationship and precision-recall curve is introduced as an approach to report the performance. Although these metric are simple and widely used, they do not take into account the ranking position (i.e. insensitive to ranking swaps above the rank cutoff), and furthermore, assumes that relevance is a binary judgement.

One of the first metrics to address the limitation of insensitivity of ranking position is AP (average precision) [Harman, 1994]:

$$AP@l = \frac{\sum_{r=1}^{l} g(r) \cdot P@r}{C_R} \tag{2.1}$$

MAP (mean AP) is one of the most widely used metric in the IR community. However, as originally conceived, AP assumes that relevance is within the binary grade.

To incorporate graded relevance and to take a more fine-grained user model into account, $nDCG@l$ was proposed [Järvelin and Kekäläinen, 2002]. By diminishing the impact of lower ranked relevant documents, $nDCG@l$ measures the performance of l by cumulating the diminished gain for each position r. A function $g(r)$ is defined to measure the gain of reading a document. The more relevant the document is, the higher the gain to the user. Finally, the metric score is normalized by an ideal ranked list l^*, obtained by ranking all relevant documents in descending order of their relevance.

$$nDCG@l = \frac{\sum_{r=1}^{l} g(r)/\log(r+1)}{\sum_{r=1}^{l} g^*(r)/\log(r+1)} \tag{2.2}$$

Generally, $g(r)$ can be chosen in different ways and in accordance with the gain function for DCG [Järvelin and Kekäläinen, 2002], it can be taken as:

$$g(r) = \frac{2^{g_r} - 1}{2^{g_{max}}} \tag{2.3}$$

where g_r is the grade of the r-th document and when the document is non-relevant (g = 0), the user gain $g(r)$ is 0, while when the document is extremely relevant (g = 4 if a 5 point scale is used), then the user gain $g(r)$ is near 1.

Other metrics have been proposed (e.g. RBP [Moffat and Zobel, 2008], ERR [Chapelle et al., 2009]); the major difference between $nDCG$ and them is the assumed user model and how $g(r)$ is defined. For RBP, where $g(r)$ indicates the degree of relevance of the document at rank r and p is a parameter that models how persistent a user is while looking through the ranked list. This measure makes similar assumptions to DCG, except the persistence parameter p models some notion of user browsing behavior, which is absent in DCG.

$$RBP@l = (1-p) \cdot \sum_{r=1}^{l} g(r) \cdot p^{r-1} \tag{2.4}$$

Although having different discount factors, both nDCG and RBP assume that the user gain $g(r)$ for inspecting a given document at rank r does not depend on the documents previously inspected. In practice, such an independence assumption does not fit well the users' observed click behaviour. As a result, ERR (expected reciprocal rank) metric is proposed to quantify the effectiveness of the ranking using the cascade user model, assuming that once a user has found the desired information, the user gain for inspecting further documents is reduced.

$$ERR@l = \sum_{r=1}^{l} \frac{1}{r} \prod_{i=1}^{r-1} (1 - g(i)) \cdot g(r) \tag{2.5}$$

Recently, there is a trend in the IR evaluation community for incorporating more user into the evaluation metrics [Clarke et al., 2013], therefore, to enable more trustworthy evaluation that correlates better with the user experience. For example, [Smucker and Clarke, 2012] proposed a *time-biased gain* (TBG) framework that explicitly calibrates the time of various (user) aspects in the search process.

$$TBG = \sum_{r=1}^{\infty} exp(-T(r)\frac{ln2}{224}) \tag{2.6}$$

where the exponential factor is the time-based decay function and $T(r)$ is the estimated time to reach rank r, computed as the time to read snippets plus the time to read clicked documents:

$$T(r) = \sum_{i=1} r - 14.4 + (0.018 l_m + 7.8) \cdot p_{click}(m) \tag{2.7}$$

This model relies on two important assumptions, namely, the linear traversal assumption (as the summation over previous ranks suggests), and that the user's reading speed is constant (the time required to read a full text is linear with respect to its length l_m as measured by the number of words).

To address the above limitations, another attempt from Sakai et al. [Sakai and Dou, 2013] proposed a unified evaluation framework (U-measure) that is free from *linear traversal* assumption. The basic idea is that U-measure first builds a trailtext which represents a concatenation of all the texts read by the user during a search session, and then computes U-measure over the trailtext, based on position-based discounting. U-measure takes the document length into account just like TBG, and has the diminishing return property. This framework is quite general and can evaluate information access other than ad-hoc retrieval (e.g. multi-document summarisation, diversified search). Recently, [Chuklin et al., 2013] proposed a common approach to convert click models into system-oriented evaluation measures. This framework enables research to propose more realistic user click models and then incorporate them into the evaluation. In addition, rather than reporting an absolute score for a system, some of evaluation attempts to capture the user variability [Carterette et al., 2011] while others elicits relative preference-based metrics [Chandar and Carterette, 2013] to measure the system performance. Furthermore, more factors (such as search engine efficiency [Wang et al., 2010]) and more user behavior data (such as mouse movements [Diaz et al., 2013, Navalpakkam et al., 2013]) are utilized for evaluation purposes.

Diversity Metrics

Recently, *diversity* has been emphasized from the IR research community and the objective is to provide a diverse ranking that could satisfy multiple information needs possibly underlying an ambiguous or multi-faceted query. This uncertainty of information need could be due to the ambiguity of the issued query (e.g. "jaguar" can be either an animal or a car brand). Even the query is not ambiguous (e.g. "google") , there could be different aspects/facets that the users are aiming for (e.g. company stock information, google search engine entry, etc.). In fact, different users or even same user in different context could aim to retrieve different aspects of the query.

Most of the diversity evaluation relies on the Cranfield paradigm, i.e. based on test collection. Like traditional IR collection, the test collection consists of documents, topics and relevance assessments. The difference lies in the fact that rather than making relevance assessments on the topic level, the assessments are made for each subtopics underlying the topic. A variety of public test collections are established in various evaluation forum (TREC, NTCIR, etc.).

To consider rewarding topical diversity in ranked lists, a set of diversity metrics have been proposed recently. The most straightforward metric for diversity evaluation is perhaps intent recall $I\text{-}rec@l$, which quantifies the amount of unique aspects of the query that are covered by the top l ranked documents. The limitation of intent recall is that it does not take into account either the ranking positions or the probability (popularity) of different aspects given the submitted query.

Other more complex metrics that aim to address the above mentioned are proposed, including α-$nDCG$ [Clarke et al., 2008], IA-$nDCG$ [Agrawal et al., 2009] and $D\#$-$nDCG$ [Sakai and Song, 2011]. α-$nDCG$ extends $nDCG$ to account for diversity by discounting the gains that accrue according to the intent (subtopic) previously encountered in the ranked list. The novelty-biased gain $NG(r)$ is defined as:

$$NG(r) = \sum_i J_i(r)(1-\alpha)^{C_n^{r-1}} \tag{2.8}$$

where $J_i(r) = 1$ if a document at rank r is relevant to the i^th intent and 0 otherwise; $C_i(r) = \sum_{k=1}^{r} J_i(r)$ is the number of documents observed within the top r results that contained the i^{th} intent. The strength of the novelty-biased discount is controlled by α.

Agrawal et al. [Agrawal et al., 2009] apply a traditional measure to each subtopic independently and then combined each value to give the expected value of the measure across all intents. This assumes that for a query q with several intents i, the probability of each intent $P(i|q)$ is available. For example, $nDCG$ for an given intent i ($nDCG_i$) is computed first, and then the intent-aware IA-$nDCG$ is computed as:

$$IA\text{-}nDCG@l = \sum_i P(i|q) nDCG_i@l \tag{2.9}$$

D-$nDCG$ [Sakai and Song, 2011], by analogy to $g(r)$ within $nDCG$, calculates a global gain $GG(r)$ at rank r given various intents:

$$GG(r) = \sum_i P(i|q) g_i(r) \tag{2.10}$$

$g_i(r)$ is the gain value for a document at rank r for intent i. Intent recall I-$rec@l$ [Zhai et al., 2003], i.e. number of intents covered by a ranked list until rank l, can be boosted with the following measure:

$$D\#\text{-}nDCG@l = \gamma I\text{-}rec@l + (1-\gamma) D\text{-}nDCG@l \tag{2.11}$$

γ controls the trade-off between relevance and diversity.

2.1.2 User-oriented Evaluation

The user-oriented evaluation can be categorized into two categories: the online evaluation (e.g. A/B testing, interleaving) that requires only behavior-based implicit feedback; and the user study (either lab-based or crowd-sourcing evaluation) that requires explicit feedback from the users.

Implicit Online Evaluation

A/B testing can be categorized into this type of evaluation approach. The basis of A/B testing [Kohavi et al., 2009] is running a bunch of single variable tests (either in sequence or in parallel): for each test only one parameter is varied from the control (the current live system). It is therefore easy to see whether varying each parameter has a positive or negative effect.

In practice, A/B testing is widely used, because A/B tests are easy to deploy, easy to understand, and it is versatile (not task-specific). For example, A/B testing approach [Kohavi et al., 2009] is frequently used by web search engines and website user experience testing. For such a test, precisely one thing is changed between the current system (A) and a proposed system (B), and a small proportion of traffic is randomly directed to the variant system, while most users use the current system. The preference of the system can then be derived from the implicit feedback derived from the usage logs, e.g. click-through rate for search engines.

Interleaving is another approach that is widely used in the search engine company to evaluate two ranking algorithms. It is a specific approach that is solely suitable for search task. Starting with the work of [Joachims, 2002], the idea of interleaving methods has become increasingly popular. The two most commonly used interleaving methods are Team-Draft Interleaving [Radlinski et al., 2008] and Balanced Interleaving [Joachims et al., 2003]. Team-Draft Interleaving can be described as follows. For each user query we build an interleaved list L whose documents are contributed by rankings A and B, the two rankings that we want to compare. This interleaved list is then shown to the user and the users' clicks are recorded. The system that contributes most of the documents clicked by the user is inferred to be the winner for the particular query-user pair; the system that wins for most such pairs is then considered to be the better system. Balanced interleaving uses a different algorithm to build an interleaved list L and a more sophisticated procedure for determining the winner.

User Study

Typical IR user studies take place in laboratory settings, where researchers are able to control the experimental environment and variables. The impact of one or more experimental variables can be isolated, and the results that are obtained are thought to be reliable due to the control of the experimental variables. Although laboratory-based user studies are useful, they are often criticised because they are too artificial, and do not represent real life search scenarios [Kelly, 2009]. In fact, users' search behaviour may have higher chance of being contaminated or biased by the experimental design or by the researchers who excessively observe users during the experiments. Besides, experiment data can only be collected for a small number of users, tasks and systems. The small size of the collected sample renders inherent population bias that may limit the generality of the studies' findings.

On the other hand, recently, crowdsourcing study has been used as an inexpensive and often efficient methodology to conduct large-scale user studies. The crowdsourcing paradigm has already been successfully used in IR for performing a number of tasks [Zuccon et al., 2013] (such as relevance assessments [Alonso et al., 2008], eliciting user's preference-based search evaluation [Arguello et al., 2011b]) Through crowdsourcing, it can capture user interactions and searching behaviors at a lower cost, with more data, and within a faster time period than traditional laboratory studies. However, it suffers from the fact that less experimental variables can be controlled during the studies and quality control [Le et al., 2010] is crucial and difficult.

2.2 Aggregated Search

We now turn to another theme of this thesis: aggregated search. As mentioned previously in Section 1, aggregated search can be viewed as one type of federated search problem that focuses on heterogeneous verticals on the web. Therefore, techniques that have been proposed for federated search can also be applied to aggregated search.

As shown in Figure 2.1, aggregated search has three main key components: vertical selection (bf VS), item selection (**IS**) and result presentation (**RP**). Vertical selection deals with deciding which verticals are (often implicitly) intended by a query. Item selection deals with selecting an optimal subset of items from each vertical to present on the aggregated page. Result presentation deals with organising and embedding the various types of vertical results on the ultimate search result page. The most common presentation strategy is to merge the results into one ranked list of blocks (i.e. *blended* presentation strategy), and is now the de facto standard for most of the commercial web search engines.

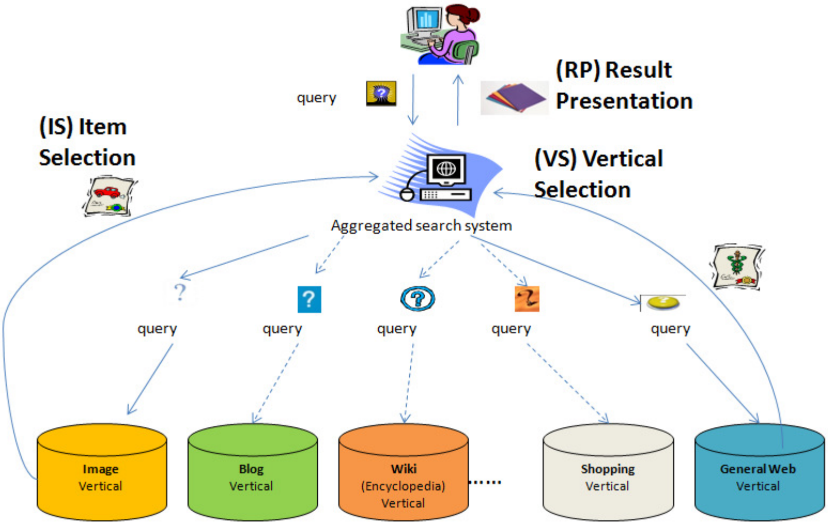


Figure 2.1: Architecture of aggregated search (three key components): (VS) vertical selection; (IS) item selection and (RP) result presentation.

We discuss each of the key component respectively from Section 2.2.1 to 2.2.3. We mainly provide a brief overview of the area and only provide details for the approaches that utilize in this thesis (e.g. for simulating aggregated search systems). A more comprehensive review of federated and aggregated search can be referred in other three survey papers published by the other domain-experts in this area [Arguello and Fernando, 2013], [Lalmas, 2011] and [Shokouhi and Si, 2011].

2.2.1 Vertical Selection

Vertical selection is one key component of aggregated search where the task is to select the relevant verticals (if any) in response to a user search query. To select the relevant vertical(s) for each submitted query, an aggregated search engine needs to know about the content of each vertical (e.g. term statistics, size, etc). This is to ensure that, for example, the query "flower" is passed to a image vertical, whereas the query "Wayne Rooney" is sent to a sport and video vertical. For this purpose, the aggregated search system keeps a representation of each of its verticals.

Vertical representation can be compared to the resource representation task in federated search, in which a number of techniques have been proposed [Shokouhi and Si, 2011]. When in a cooperative environment where all the collection statistics can be obtained, normally the collection statistics of each vertical are utilized for the vertical representation. However, there are also limitations underlying this. Firstly, it is not always the case that all the verticals can be accessed (e.g. in a uncooperative environment). Secondly, even all the verticals can be accessed, to use the whole collection as the representation, might be not feasible due to efficiency issue to access such large term statistics for all the verticals.

Therefore, other techniques have been proposed while the most well-known strategy is called query-based sampling approach [Callan and Connell, 2001]. The main idea is to iteratively issue the query to a given vertical search engine and then utilize the a small sample of retrieved results (documents) to form the vertical representation. The major differences among the variety of approaches proposed is how the queries are sampled (e.g. from different resources, for instance, the retrieved documents [Callan and Connell, 2001] or external resources such as the query-log [Shokouhi et al., 2007] or Wikipedia [Arguello et al., 2009]). In addition, for a given resource to sample documents from, the way to sample the queries could be different as well (e.g. random sampling or popularity-based sampling [Callan and Connell, 2001]).

Given the vertical representation, vertical selection deals with ranking and selecting the most relevant vertical for a given query. Vertical selection can be compared to the resource selection task that are well investigated in federated search, where the most common approach is to rank resources based on the similarity between the query and the resource representation. It has been shown in previous work [Si and Callan, 2003a] that it is not practical to issue the query to every resources to obtain relevant contents since the relevance of information conforms to a skewed distribution. Therefore, the objective of resource selection is to select a subset of $R_k = \{r_1, ...r_k\}$ resources that potentially maintain most relevant documents for the query q from all the resources $R_n = \{r_1, ...r_n\}$ available. Generally, in federated search [Shokouhi and Si, 2011], it is assumed in most prior work that k is pre-defined and therefore, resource selection can be solved as a resource ranking problem, i.e. given the query q, selecting the top-k ranked resources that are estimated most relevant.

Therefore, the major differences among different available resource selection approach are the way they estimate the relevance of the resource given the query. Current resource selection approaches can be categorized into three categories:

- Query-Vertical Similarity Approach

- Relevance Distribution Estimation Approach
- Machine Learning Based Approach

While these approaches all derive evidence from the same source (sampled vertical representation), they model different aspects of the sources under consideration. For example, CORI, Clarity and GAVG model the similarity between the query and the source, whereas ReDDE, CRCS(l) and CRCS(e) model the collection's average document score in a full-dataset retrieval (all sources together). We discuss each types of those approaches in more details in the following sections. Note that those discussed approaches (especially the ones with formulas) will be utilized in our experiments in the remaining of the thesis for simulating aggregated search systems.

Query-Vertical Similarity Approach

The query-vertical similarity approach is the most common approach for resource selection since it is directly adapted from traditional single collection retrieval. This is intuitive and the basic idea is to mainly treat each vertical as a document and retrieve from there. The main approaches in this category include CORI [Callan et al., 1995], Clarity [Cronen-Townsend et al., 2002].

CORI

CORI adapts INQUERY's inference net document ranking approach to collection. Here, all statistics are derived from sampled documents rather than the full collection. The CORI resource selection algorithm [Callan et al., 1995] calculates belief scores of individual collections by utilizing a Bayesian inference network model with an adapted Okapi term frequency normalization formula

$$CORI_q(v_i) = \prod_{w \in q} \left\{ b + (1-b) \times \frac{df_w^{v_i}}{df_w^{v_i} + 50 + 150 \times cw_{v_i}/avg\_cw} \times \frac{log(\frac{|C|+0.5}{cf_w})}{log(|C|+1.0)} \right\} \tag{2.12}$$

where $df_w^{v_i}$ is the document frequency of query term w in the vertical collection C_i, cf_w is the number of collections containing query term w, $|C|$ is the number of vertical collections, cw_{v_i} is the number of terms in C_i, avg_cw is the average number of words per collection, and b is the default belief.

Clarity

Clarity is a retrieval effectiveness prediction algorithm [Cronen-Townsend et al., 2002] that measures the similarity between the language of the top ranked documents and the language of the vertical collection, estimated using the Kullback-Leibler divergence between the query θ_q and the vertical collection language model θ_{v_i}.

$$Clarity_q(v_i) = \sum_{w \in v_i} P(w|\theta_q) log_2 \frac{P(w|\theta_q)}{P(w|\theta_{v_i})} \tag{2.13}$$

where $P(w|\theta_q)$ and $P(w|\theta_{v_i})$ are the query and vertical v_i collection language models, respectively. The query language model is normally estimated using the top 100 documents retrieved. Claritys assumption is that in an effective retrieval the top ranked documents will use language that is distinguished from topic-general language derived from the entire index of the vertical collection.

Other approaches, such as the one proposed in [Si and Callan, 2002], rank collections based on other similarity measure (e.g. the Kullback-Leibler divergence) between the query and vertical collection language model. These query-vertical similarity models have the advantage of being straight-forward adaptations of well-studied document ranking techniques. However, because they do not distinguish between documents in the vertical collection, the vertical collection language model is dominated by the larger documents. This is a potential disadvantage if we care about relevance at the document level. Also, they compare the text in the query with the text in the entire vertical collection, making no distinction between documents that are related and unrelated to the query. Because of this, this query-vertical similarity approach generally favors collections with a high proportion of relevant-to-non-relevant documents. They may favor a small topically focused vertical collection (related to the query) over a larger more topically diverse vertical collection that contains more relevant documents.

Relevance Distribution Estimation Approach

Rather than treating the vertical as a document, another category of approach aims to model in a more refined way by estimating the distribution of relevant documents within each vertical.

ReDDE

The relevant document distribution estimation (***ReDDE***) resource selection algorithm [Si and Callan, 2003a] was designed to select a small number of collections with the largest number of relevant documents. To achieve this goal, ReDDE explicitly estimates the distribution of relevant documents across all the resource collections and ranks those resources accordingly. ***ReDDE*** scores a resource collection based on its expected number documents relevant to the query while it derives this expectation from a retrieval of an index that combines documents sampled from every vertical collection. Given this retrieval, ***ReDDE*** accumulates a collection score $ReDDE_q(v_i)$ from its document scores $P(q|\theta_d)$, taking into account the difference between the size of the original collection N^{v_i} and a sampled set size N^{samp}.

$$ReDDE_q(v_i) = \frac{N^{v_i}}{N^{samp}} \sum_{d \in topm} I(d \in v_i) P(q|\theta_d) \tag{2.14}$$

where $I(.)$ is a indicator function. Different variants of the ReDDE algorithms have emerged, which weight top-ranked documents and estimate the probability of relevance in different ways.

CRCS

Like *ReDDE*, *CRCS* is proposed [Shokouhi, 2007] and the idea is to issue the query to a centralized sample index and score a resource collection according to an accumulation of a more refined estimation of document score. Specifically, the document score for *CRCS(l)* and *CRCS(e)* are estimated by a linear or a negative exponential weighting according to its presented position respectively.

$$CRCS(l)_q(v_i) = \frac{N^{v_i}}{N^{samp}} \sum_{d_j \in topm} (m - j) \tag{2.15}$$

$$CRCS(e)_q(v_i) = \frac{N^{v_i}}{N^{samp}} \sum_{d_j \in topm} \alpha \cdot exp(-\beta \cdot j) \tag{2.16}$$

where $\alpha = 1.2$ and $\beta = 2.8$ in our setting. In particular, the weight represents the impact of a given document d at the jth position of the results returned by the centralized index of all sampled documents. Parameter m specifies the number of top-ranked documents that are considered as relevant, and was set to 50 by [Shokouhi, 2007]. In CRCS(l), the impact of documents decreases linearly according to their ranks while in CRCS(e), a variant in which the importance of documents drops exponentially. The exponential version of CRCS(e) has been reported in [Thomas, 2008] to produce slightly better results compared to the linear form.

Geometric Average

Geometric Avearge (*GAVG*) approach proposed in [Seo and Croft, 2008] issues the query to a centralized sample index, one that combines document samples from every vertical, and scores vertical v_i by the geometric average query likelihood from its top m sampled documents.

$$GAVG_q(v_i) = \left(\prod_{d \in topm} P(q|\theta_d) \right)^{\frac{1}{m}} \tag{2.17}$$

Recently, other more complex approaches have been proposed along with this line [Markov et al., 2013b, Hong and Si, 2013]. For example, a perturbation approach has been proposed in order to derive risk-aware resource selection [Markov et al., 2013b]. The main idea is to reduce the uncertainty of resource selection by producing a set of rankings from query perturbation, retrieval system perturbation and ranking perturbation. Then by aggregating over the different rankings, better estimated resource selection scores with less uncertainty is obtained. In addition, recently, the diversity of results is considered [Hong and Si, 2013] in federated search and this has been incorporated in the resource selection. The basic idea is not to only select resources with the largest number of relevant documents, but also to select the resources that cover multiple different aspects underlying the issued queries.

Machine Learning Based Approach

Since prior work shows that no single source of evidence can be used to predict that a particular vertical is relevant to a query [Arguello et al., 2009], therefore, a variety of work [Arguello et al., 2009, 2010, Li et al., 2008] have been proposed to model vertical selection as a machine learning approach (i.e. classification task), and aim to use machine learning to combine multiple types of predictive evidence as features. There are mainly three types of features [Arguello and Fernando, 2013]: query features, vertical features and vertical-query features.

Query features are generated from the query itself and not from any resource associated with a candidate vertical. The query features are normally generated from the query string and not from any vertical representation, for example a query string with the presence of a particular term (e.g. "news", "weather") [Arguello et al., 2009, Li et al., 2008] or a particular named entity (e.g. person) [Arguello et al., 2009].

Vertical features are generated from the vertical, independent of the query. This is due to the fact that some verticals are more popular than others either because they satisfy a wider range of information needs (e.g. news) or because they satisfy more frequently occurring information needs (e.g. weather). This can be analyzed and generated by the vertical representation using the offline fashion and have been demonstrated to be useful in the news vertical [Diaz, 2009].

The third type of feature *vertical-query feature* is mostly relevant to the resource selection approaches we have discussed (such as CORI and ReDDE). Vertical-query features aim to measure relationships between the vertical and the query and are therefore unique to the vertical-query pair. Basically any single resource selection approach can be utilized as one feature for the vertical-query features. According to [Arguello and Fernando, 2013], vertical-query features can be classified into pre-retrieval, post-retrieval, and post-presentation features. Since in this thesis, we mainly focus on pre-retrieval feature where most of the resource selection approaches can be categorized into, therefore, we will not discuss the post-retrieval and post-presentation features. More details can be referred in [Arguello and Fernando, 2013].

2.2.2 Item Selection

Item selection deals with selecting an optimal subset of items for the final presentation from all the documents retrieved from the selected verticals (in addition to the default "general web" results). Most of current work assumes utilizing the ranking of each vertical and selecting a fixed number of documents for each selected vertical (e.g. five images or three news articles) for presentation in the vertical block in the ultimate SERP. This key component of aggregated search has not been much investigated in the research community.

However, different ranking functions generally exist for different verticals. This is due to that given different types of information, different types of features and their corresponding importance (weights) are different from each other for ranking documents in different verticals. For example, the features utilized in the image ranking approach (e.g. visual saliency of the image) is different from the one used to ranking news (e.g. temporal recency). In addition, even for the same vertical, the effectiveness of each search engine could vary significantly [Si and Callan, 2005]. Therefore, this should also have to be considered whether to re-rank the retrieved documents from a vertical or not, in order to improve the document ranking in the selected vertical. Furthermore, it is not clear what is the optimal number of documents to be presented for each vertical.

Recently, a new evaluation measure [Lv et al., 2014] have been proposed to incorporate evaluating a vertical ranking with different number of presented results. Recent research [Arguello and Capra, 2012, Arguello et al., 2013] also shows that the coherence of the vertical results also have impacts on user's behavior on general web results. This implies that relevance is not the sole criteria of selection of items to be presented while the relationship among items across verticals should have also been taken into consideration.

Therefore, the above research all demonstrate that rather than presenting each vertical ranking independently, it might be important for the aggregated search system to select an optimal subset of items. Although in this thesis, this is not the main focus, we degrade differently on the ranking function for different verticals in our simulation in the aggregated search systems and evaluate a large set of different (albeit simple) item selection strategies.

2.2.3 Result Presentation

Vertical presentation refers to deciding precisely where to place relevant verticals on the search results page. The presentation strategies utilized are different between traditional homogeneous search federation and the heterogeneous aggregated search scenario. In the homogeneous federation, normally retrieved results (ranked lists) from different selected resources are merged into one unified ranked list. This is generally as known as result merging task in traditional federated search. However, in heterogeneous aggregated web search, the results retrieved from the selected verticals can be presented in a variety of presentation strategies we have mentioned (e.g. blended, non-linear blended or tabbed) in Chapter 1. In this work, we constrain our discussion to blended presentation strategy for heterogeneous aggregated web search, i.e. a ranked list of web documents and interleaved with vertical results, due to its popularity in the research and commercial community. We discuss the different research efforts in homogeneous result merging and heterogeneous result presentation respectively as below.

Homogeneous Result Merging

In result merging [Shokouhi and Zobel, 2009], the federated search system receives the top-ranked results of selected collections and orders them in a single list for presentation to the user. The main task in result merging is to compute comparable scores for documents returned by different resource collections. Since this is not main focus of the thesis, we only briefly discuss the most common strategies utilized. The most common result merging is CORI [Callan et al., 1995] and it was proposed in early 1995. The CORI result merging approach is a linear combination of the collection selection scores and the document scores returned by collections. CORI uses a simple heuristic formula to normalize collection-specific document scores. Later, the parameter of the heuristic combination of CORI approach has been re-visited [Markov et al., 2013a] in order for a more principled and better estimation.

Later, SSL approach is proposed [Si and Callan, 2003b] and this is a semi-supervised learning method for result merging. SSL trains a regression model for each collection that maps document scores into their global merging scores. The basic idea is to run the query against a central sampled index that index a subsampled documents from all available resources (verticals), and compare the centralized scores of such overlap sampled documents with the scores reported by vertical collections to compute the merging scores. Other approaches, such as SAFE [Shokouhi and Zobel, 2009], has been recently proposed and the idea is to use the scores of all documents in agglomeration of all the collection samples, and generates a statistical fit to estimate scores.

Heterogeneous Result Presentation

As we have mentioned, result presentation research in aggregated search has been mainly investigated in the context of blended presentation strategies. This can be casted as a block-ranking problem and has been investigated by utilizing either the offline setting [Arguello et al., 2011a] or the online setting [Jie et al., 2013, Ponnuswami et al., 2011b].

[Ponnuswami et al., 2011b] is the first work that considers multiple verticals to be presented at different positions on the SERP. More specifically, this work considered three slotting positions: top (above the Web results), middle (between Web results 3-4), and bottom (below the last Web result). Their proposed framework treats this problem as a classification problem and combines vertical-specific binary classifiers by using thresholding parameters for each position. Then each selected vertical is assigned to the top, middle, or bottom position based on its classifier's prediction confidence value.

In [Arguello et al., 2011a], in addition to the supervised classification framework proposed before, a variety of learning to rank approaches have been developed to address this block-ranking problem. In particular, the authors utilize a variety of features that are generally correlate with the features that we have discussed in the vertical selection task and investigate in this presentation optimization task. After extensive experiments, they conclude that the best overall performance is obtained by casting block-ranking as a learning-to-rank problem.

Recently, [Jie et al., 2013] have proposed an unified framework for the search federation problem and they model the search federation as a contextual bandit problem to optimize the result presentation of vertical blocks. Basically, the system leverages implicit user feedback to explore and exploit on users' preference on different vertical and corresponding vertical presentation positions, and then learn the result presentation model which maximizes the total reward for the users.

In our work for simulating result presentation strategies, we employ a similar strategies as in [Ponnuswami et al., 2011b] that we utilize a multi-label classification model and the corresponding learned thresholds for different presentation positions to determine the ultimate embedding positions for the vertical blocks.

2.2.4 Evaluation

In general, the goal of evaluation is to facilitate the objective comparison between different aggregated search algorithms and ultimately the user experience on different aggregated search pages. In this section, we only provide a brief overview of previous work and details of these can be referred in the corresponding most relevant chapters.

Various works evaluating one component of an aggregated search system in isolation exist. Vertical selection in aggregated search has been studied in [Arguello et al., 2009, 2010, Li et al., 2008]. Much of this research aims to measure the quality of the set of selected verticals, compared with an annotated set obtained by collecting manual labels from assessors [Arguello et al., 2009, 2010] or derived from user interaction data [Li et al., 2008, Ponnuswami et al., 2011b]. The annotation can be binary [Arguello et al., 2009] or graded [Ponnuswami et al., 2011b]. The quality of a particular vertical selection approach is mostly evaluated with standard measures of precision and recall using the binary annotated set. Within the context of resource selection in the homogeneous federated search, metrics for evaluating resource selection methods are usually recall-based [Shokouhi and Si, 2011], i.e. resource selection techniques are compared according to the number of relevant documents available in selected resources.

Recent attempts to evaluate the utility of the whole aggregated search page [Arguello et al., 2011b, Bailey et al., 2010] consider the three key components of aggregated search (VS, IS, RP) together. For evaluating the merged ranked lists in traditional homogeneous federation, different approaches are usually compared according to the number of relevant documents in the final merged results based on a test collection.

For evaluating the entire aggregated search system, the evaluation falls under three broad categories: test collection evaluation, on-line evaluation, and user study evaluation. For example, [Ponnuswami et al., 2011b] and [Jie et al., 2013] falls into the online evaluation approach and they evaluate the utility of a page based on a user engagement metric (CTR). This evaluation framework requires large-scale user interaction data, which may not always be available. If the goal is to fine-tune an existing system, it may not be possible to conduct an on-line evaluation for every combination of parameter values. To address this limitation, a few recent studies [Ponnuswami et al., 2011a] have investigated methods for collecting on-line user-interaction data once and using this data to perform multiple rounds of offline testing.

Several research work [Arguello and Capra, 2012, Bailey et al., 2010, Bron et al., 2013, Sushmita et al., 2010] fall into the user study evaluation category. [Bailey et al., 2010] evaluate the utility of the page by asking annotators to make assessments based on a number of criteria (e.g. relevance, diversity) on the ultimate SERP. [Sushmita et al., 2010] have investigated the impact of aggregated search interfaces (different vertical orientation and presentation strategies) to user search behavior (click-through rate). They found that users click more on verticals that are relevant to the task, verticals that are shown higher in ranking and verticals that are more visually salient. Both [Arguello et al., 2012] and [Bron et al., 2013] investigated the effect of task complexity on users' demand for vertical content and they found that during more complex tasks, users exhibit a greater demand for content that is more diverse (i.e. multiple verticals). Recently, [Arguello and Capra, 2012, Arguello et al., 2013] investigated result coherence, the effect of the query-senses represented in the blended vertical results on user interaction with the web results. They found that given an ambiguous query (e.g., "jaguar"), user interaction with the web results is greater when the query-senses represented in the vertical results are consistent with the intended query-sense. They call this "spill-over" effect and this is especially true for some verticals (images, shopping, video), but not others (news).

With respect to test collection evaluation, [Arguello et al., 2011b] collects preferences on block pairs from users and measures the page quality by calculating the distance between the page in question and the ideal (reference) page; the shorter the distance, the better the page. One advantage is that any possible combination of vertical blocks that form an aggregated page can be tested, from a block-oriented point of view (without regard to item selection). Others [Santos et al., 2011] have proposed an aggregated search metric that captures both vertical diversity and topical diversity.

Part II

On the Assessments of Aggregated Search

3

Assessing and Understanding Vertical Relevance

We start our Part II of this thesis, aiming to understand better on the assessments for Cranfield paradigm for aggregated search. In this chapter, we start our investigations by studying whether different underlying assumptions made for vertical relevance affects a users perception of the relevance of verticals. This chapter addresses our research questions **RQ 1** to **RQ 4**, as specified in Section 1.2.1. We present a formal analysis and a set of extensive user studies to study this.

Current approaches that evaluate the effectiveness of aggregated search (AS) systems are based on rewarding systems that return highly *relevant* verticals for a given query, where this *relevance* is assessed under different assumptions. Despite the relative success of these evaluation methodologies, the definition of the *relevance* of a vertical, given a query, remains unclear. Although some prior work [Arguello et al., 2009, 2011b, Ponnuswami et al., 2011b, Zhou et al., 2012a] exist on assess vertical relevance, either by explicitly ask annotators or implicit derive from use behavior data, it is difficult to evaluate or compare those systems without fully understanding the relationship between those underlying assumptions. Understanding these underlying assumptions is important and this is because a different annotation set (i.e. gold standard) will affect the metrics that inform us about the performance of different VS systems. Our main aim is to provide insights into various aspects of query vertical *relevance* and allow us to explain in more depth as well as questioning the evaluation results published in the literature.

In this thesis, the *relevance* of a vertical for a given query refers to the perceived usefulness of the vertical on a SERP. The underlying assumptions made when assessing the relevance of verticals may have a major effect on the evaluation of a SERP. Consider a user who issues the query “yoga poses” to an AS system that has access to five verticals (‘news’, ‘image’, ‘video’, ‘shopping’ and ‘blog’). Prior to viewing the aggregated results, the user may believe that both the *‘image’* and *‘video’* vertical might provide more relevant results. If such a pre-retrieval evaluation is conducted, the user might annotate those two verticals as relevant. Conversely, a user who viewed the retrieved results from each vertical might conclude that *‘video’* and *‘blog’* provided the most relevant results. This may be due to the presence of a blog article that comprehensively describes yoga poses and a highly ranked *‘video’* vertical that contains similar information to an *‘image’* vertical that appears lower down the ranking. In this case the *‘image’* vertical may seem to provide redundant information. These scenarios give us some insight into the complexity of defining the relevance of verticals.

A total of more than 20,000 assessments on 44 search tasks across 11 verticals are collected through Amazon Mechanical Turk and subsequently analysed. Firstly, we aim to compare vertical relevance assessments assessed at two different search stages:

> **RQ 1**: Are there any differences between the assessments made by users from a pre-retrieval user-need perspective (viewing only vertical labels prior to seeing the final SERP) and the assessments made by users from a post-retrieval user perspective (viewing the vertical results in the final SERP)?

Pre-retrieval vertical relevance assessments may differ to post-retrieval ones. This could be due to serendipity (finding a surprisingly excellent result from a specific vertical) or to a poorly designed vertical (a poor ranking function within the vertical). We found that both *orientation* (pre-retrieval user need) and *topical relevance* (post-retrieval topical relevance) correlates significantly with the post-retrieval search results utility. The impact of orientation is comparatively more significant (moderate) than topical relevance (low). In addition, there is an aesthetic bias to a user's perception of search results utility.

Secondly, when aggregated search systems present vertical items embedded within *'general web'* results, it is not clear whether using *'general web'* as a reference for deciding the vertical *relevance* is an appropriate strategy. Therefore, we aim to compare vertical relevance assessments at their dependency assumptions. In addition, it is possible that making independent vertical relevance assessments does not reflect the characteristics of aggregated search, such as avoiding redundancy (an *'image'* result containing information already presented in a *'video'* result). Therefore, we aim to study the impact of contextual information (in another vertical) to the user's assessments:

> **RQ 2**: When using "general web" results as a reference for making vertical relevance assessments, are these assessments able to predict the users' pairwise preference between any two verticals?

> **RQ 3**: Does the context (results returned from other verticals) in which vertical results are presented affect a user's perception of the relevance of the vertical of interest?

We conclude that "general web" results can be served as a reference for deciding vertical relevance and it is effective from the utility-effort perspective in collection assessments. In addition, the context of other verticals has significant impact on the relevance of a vertical.

Finally, it is not clear whether user's preference assessment could be used to infer the positioning of vertical results. This is important to understand this since we aim to utiilze this assessments to evaluate result presentation of aggregated search as well. Therefore, we aim to answer:

> **RQ 4**: Is the vertical preference information provided by a population of users able to predict the "perfect" embedding position of a vertical?

We found that it is possible to employ a number of binary assessments to predict multi-grade assessments and the correlation of the derived optimal pages is significant (moderate). Using a larger number of assessments contributes to more accurate estimation of multi-grade assessments.

The remainder of this chapter is organized as follows. In Section 3.1, we formally outline the problem of vertical selection assessment. Section 3.2 outlines our experimental design, whereas in Section 3.3 we present and analyse our results. We conclude the chapter in Section 3.4 by summarising all the findings, discussing the implications of our results and pointing out limitations.

3.1 Vertical Relevance Assessments

We start by defining the process involved in collecting vertical relevance assessments. Second, we enumerate the various components within aggregated search that affect vertical relevance assessments and outline their relationships. Thirdly, we review various approaches that derive vertical relevance from the collected assessments. We then present an analysis of the assumptions made in previous work and discuss how they can affect the evaluation of aggregated search systems. We end this section with a summary.

3.1.1 Assessment Process

Before formally defining the vertical relevance assessment process, we first list the assumptions made for a SERP P. Given a set of verticals $V = \{v_1, v_2, ...v_n\}$, a SERP P can be denoted as $V_p = \{v_{p1}, v_{p2}, ..., v_{pn}\}$ where each v_{pi} indicates the position of the vertical block v_i on the page. For consistency with existing work [Ponnuswami et al., 2011b], we assume four positions in which verticals can be embedded into the 'general web' results: Top of Page (ToP), Middle of Page (MoP), Bottom of Page (BoP), or Not Shown (NS). When we are only interested in a binary scenario (shown or not), it is assumed that it is best to present the vertical at ToP. Note that in V_p, multiple verticals can have the same grade (e.g. two verticals can be simultaneously shown at ToP).

Given a vertical set $V = \{v_1, v_2, ...v_n\}$, the vertical relevance I_t for a search task t is represented by a weighted vector $I_t = \{i_1, i_2, ...i_n\}$, where each value i_k indicates the importance of vertical v_k to search task t. Commonly, I_t is a binary vector [Arguello et al., 2009, Zhou et al., 2012a], where each element indicates whether or not the vertical is *relevant* given the search task. When denoting the best position in which to embed the vertical items in the SERP (ToP, MoP, BoP, NS), a weighted vector I_t can be used [Ponnuswami et al., 2011b, Arguello et al., 2011b]. By assigning diminishing weight according to the embedding position,[1] each weight $i_j \in I_t$ of vertical v_j is represented by the corresponding assigned weight of its perfect embedding position.

To generate I_t, user studies must be conducted asking an assessor $u_j \in U = \{u_1, u_2, ...u_m\}$ to make decisions $A_j = \{a_{j1}, a_{j2}, ...a_{jl}\}$ over all verticals V. There are generally two types of assessment a_{jk}: absolute assessments ("what is the quality of v_i") and preference-based assessment ("does v_i present better information than v_j"). As AS is concerned with presenting vertical results integrated within 'general web' results, preference assessments [Ponnuswami et al., 2011b, Zhou et al., 2012a, Arguello et al., 2009,

[1] The higher the position, the larger the weight is, i.e. for the four embedding positions used in our work, weight(ToP) $>$ weight(MoP) $>$ weight(BoP) $>$ weight(NS).

2011b, Zhou et al., 2012b,c] have been more widely used. The number of pair-wise assessments l the assessor u_j needs to make for A_j is a matter for research, and may be restricted by the budget of a particular study. Regardless, for each pair-wise preference assessment a_{jk}, there are various factors that influence assessors' decisions. We discuss these in Section 2.2. Ultimately, an $m \times l$ matrix M_t containing all assessments from all users in U for search task t is obtained. A conflation method to derive the final vertical relevance vector I_t from the matrix M_t is used. Different methods have been used to derive this final vector, which we review in Section 2.3.

After I_t is obtained, an aggregated search page P can be evaluated based on this information. Given I_t, we can evaluate the SERP P based on how V_p correlates with I_t. Various metrics can be employed to achieve this. Precision, recall and the f-measure have been used when I_t is treated as a binary decision [Arguello et al., 2009, Zhou et al., 2012a]. Recently, risk has been considered and incorporated into risk-aware VS metrics [Zhou et al., 2012b]. When allowing multiple embedding positions within a SERP, the distance between V_p and a perfect page $V_p^{Perfect}$ derived from I_t can be used [Arguello et al., 2011b]. The further the distance from the perfect page, the worse the performance of the system that generated that SERP P.

3.1.2 Making Preference Assessments

This section reviews previous work on making preference assessments for evaluating vertical relevance.

Dependency of Relevance

Current work on determining the preference assessments A can be classified into two categories: *anchor-based* and *inter-dependent* approaches. The former assumes that the quality of the anchoring 'general web' results serve as a reference criteria for deciding vertical relevance (whether an assessor believes the vertical results will improve the SERP when added to the 'general web' results). This is achieved by asking assessors to assess each vertical v_i individually, in an independent pair-wise fashion against the 'general web' reference page. A number of works [Arguello et al., 2009, Ponnuswami et al., 2011b, Zhou et al., 2012a] follow this approach. *Inter-dependent* approaches assume that the quality of verticals is relative and dependent on each other. These approaches gather pair-wise preference data over any, and many, possible pairs of verticals v including the '*general web*' w. [Arguello et al., 2011b] fits into this category. For *anchor-based approaches*, the number of assessments to be made per assessor, l, equals to the number of verticals n. For *inter-dependent approaches*, l will often be much greater than n (e.g. $\frac{1}{2} \cdot (n+1) \cdot n$ in [Arguello et al., 2011b]).

Table 3.1: Summary of Vertical Relevance Assumptions Made in Previous Works.

Work	Relevance Dependency		Influencing Factors			Assessment		# Assessors
	Inter-dependent	Anchor	Result Quality	Orientation	Aesthetic	Binary	Graded	
Federated Search [Shokouhi and Si, 2011]	✓		✓			✓		1
Zhou et al. [Zhou et al., 2012a]		✓		✓			✓	4
Ponnuswami et al. [Ponnuswami et al., 2011b]		✓	✓	✓	✓		✓	1
Arguello et al. [Arguello et al., 2011b]	✓		✓	✓	✓	✓		4

Influencing Factors

Various factors can affect a user u_j when assessing a_{jk}, with respect to a specific vertical result v_k:

- **(Result Quality)** the quality of the retrieved results from vertical v_k.

- **(Orientation)** a user's (u_j) orientation (or preference) to information from a vertical v_k.

- **(Aesthetic)** the aesthetic nature of a vertical v_k.

The *result quality* of the retrieved items from a specific vertical depend on both the contents of the vertical v_k and the ranking function of the vertical v_k. For a given search task t, the more topically relevant items contained in the vertical v_k collection, the better the results are likely to be. More importantly the higher the relevant items are ranked within the vertical, the better the *result quality* is. Either a vertical v_k collection with very few relevant items or a poor ranking function can degrade the user's perception of the quality of the vertical v_k retrieved results.

A user's *orientation* to a vertical v_k reflects the user's (u_j) own perception of the usefulness (utility) of the vertical to the search task t. The user may have his or her own personalised preference over different verticals. As pointed out in [Arguello et al., 2009, Sushmita et al., 2010], it is not only *result quality* that satisfies a user's need, but items from different verticals also satisfy a user's need differently. It is the *type of information* that affects the user's perception of usefulness (i.e. orientation) for an information need.

Vertical *aesthetics* represents the aesthetic nature of the vertical v_k retrieved results. For example, it has been demonstrated in [Arguello et al., 2011b, Sushmita et al., 2010] that the visually attractive nature of image results tends to increase users engagement on a SERP, compared to those that do not contain images.

3.1.3 Deriving Relevance from Assessments

The anchor-based and inter-dependent based approaches use different strategies for deriving vertical relevance (I_t) from the assessments (M_t) for a search task t. For *anchor-based approaches*, most of previous work [Arguello et al., 2009, Zhou et al., 2012a] rank all the verticals of interest based on the percentage of assessors' preference over a 'general web' anchor. Therefore, a majority preference for a particular vertical leads to the most *relevant* vertical for a specific search task. For *inter-dependent approaches*, the Schulze voting method [Arguello et al., 2011b, Schulze, 2011] is the most widely used. For two verticals v_i and v_j, if more assessors preferred v_i over v_j than vice versa, then we say that, v_i directly beats v_j. A beatpath from v_i to v_j can be either a direct or an indirect defeat. The strength of an indirect beatpath is the number of votes associated with its weakest direct defeat. Finally, v_i defeats v_j if the strongest (direct or indirect) beatpath from v_i to v_j is stronger than the one from v_j to v_i. All verticals of interest are then ranked by their number of defeats.

3.1.4 Prior Work

When collecting an assessment a_{jk}, current work makes a number of different assumptions (dependency of relevance, influencing factors) to guide the assessments. Based on the assumptions made, they show the corresponding information to the user for them to make assessments. We formally review and summarize the underlying assumptions made in a number of studies. A short summary is given in Table 1.

Traditionally, in federated search [Hawking and Thomas, 2005, Shokouhi and Si, 2011] (often known as *distributed* information retrieval), vertical relevance I_t is assumed to solely depend on *result quality*, which is determined by the summation of the number of topically relevant items within a vertical collection. The more topically relevant items the vertical collection contained, the better the given vertical is assumed to be. When evaluating a SERP P, the quality of the page is determined by evaluating the topical relevance of the items returned (and merged from various verticals), based on traditional information retrieval metrics (e.g. precision, MAP). This type of evaluation is heavily focused on topical relevance.

In aggregated search, for example, [Zhou et al., 2012a] assumed that only vertical *orientation* contributes to the usefulness of the page. Therein, the assessors are asked to use the '*general web*' results as an anchor to assess the usefulness of a given vertical (by only showing the vertical label). Without viewing the retrieved results or the vertical collection, only when the assessor thinks that the vertical can potentially provide more appropriate results than the '*general web*', would he/she label it as *relevant*. In that research, four assessors are asked for assessments for each vertical. The vertical relevance I_t is determined in a binary manner (ToP or None), by using a basic assessor preference thresholding approach (e.g. if 75% of the assessors prefer v_i over w, then we label v_i as "ToP", otherwise we label it as "NS"). Finally, VS evaluation is based on the f-measure.

In [Arguello et al., 2011b], although not stated explicitly, it is assumed that the usefulness of the vertical v_k is determined by a combination of *result quality*, *orientation* and *aesthetics*. While viewing results retrieved from each vertical collection using a ranking function unique to the vertical, the assessors are asked to state the preference between any two verticals from $V \bigcup \{w\}$. Four assessors are used for assessing each pair. Different from [Zhou et al., 2012a], which uses '*general web*' results as an anchor, the assessments are made between any v_i and v_j pairs and a voting strategy is used to determine I_t, i.e. the perfect position of the vertical to be presented. The quality of the page is then measured by calculating the distance to a reference page (a "perfect" AS page).

In [Arguello et al., 2011b, Ponnuswami et al., 2011b], a vertical relevance is assessed by presenting the search page with the web results and vertical results separately. In [Ponnuswami et al., 2011b], the assessors are asked to rank the vertical relevance on a scale of 0 to 3, indicating whether it should be shown at BoP, MoP or ToP. Only one assessor is used. The differences between [Arguello et al., 2011b] and [Ponnuswami et al., 2011b] is that, instead of voting across all verticals, the '*general web*' retrieved results are used as an anchor to determine the vertical importance.

3.1.5 Summary of Aims

Given the more formal treatment of the task of aggregated search described in this section, these research questions can be stated as follows:

- **RQ 1** deals with comparing the user perspective ([Zhou et al., 2012a] and [Ponnuswami et al., 2011b] (binary assessment variant)) during the assessment stage (obtaining a_{jk}). When asking assessors to make a_{jk}, are there any differences between the assessments made by only considering *orientation* (pre-retrieval perspective), and the ones that consider a combination of *result quality*, *orientation* and *aesthetics* (post-retrieval perspective)?

- **RQ 2** and **RQ 3** are concerned with comparing the anchor-based approach with an inter-dependent approach ([Ponnuswami et al., 2011b] (binary assessment variant) and [Arguello et al., 2011b]) during the collection of all assessments A with respect to v_k. We also examine whether the context of other verticals can affect the relevance of the vertical of interest.

- **RQ 4** deals with the positioning of vertical results. When asking a set of assessors to make assessments a_{jk} using a binary decision (ToP and NS), is it possible to use the fraction of assessors' preference assessments M_t to derive an accurate graded vertical relevance I_t to indicate the best position for embedding the vertical results (ToP, MoP, BoP and NS)?

3.2 Experimental Design

This section introduces the methodology for conducting our users studies, followed by a detailed design of each study.

3.2.1 Methodology

We conducted three studies that follow a similar protocol. All studies consisted of subjects that pair-wisely assessed the quality of two result sets for a series of search tasks. All studies have a similar objective, to investigate the correlation between the vertical *relevance* derived when using one assessment assumption to the vertical *relevance* derived under another assumption.

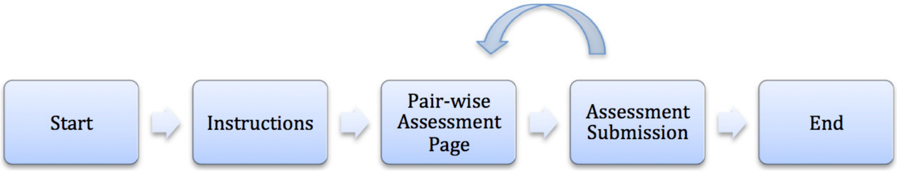


Figure 3.1: Flow Diagram Description of Experimental Protocol for Studies 1 to 3 in Understanding Vertical Relevance Assessments.

Protocol

The three studies follow a similar protocol shown in Figure 3.1. Subjects were given access to an assessment page that consists of a task description, a search task and two search results (tiles), and were asked to make pair-wise preference assessments. Prior to each study, the subjects were presented with a brief instruction, summarizing the experimental protocol and the assessment criteria. They were told to imagine they were performing a natural information search task. Given two search result sets originating from two search engines, the subjects were told to select the result set that would best satisfy the search task. The subjects were then presented with an Assessment Page (ASP) (a screenshot of an ASP is shown in the middle of Figure 3.2). The experimental manipulation was controlled via each ASP, as discussed in Section 3.1.3.

Following a search query (e.g. "living in India") shown at the top of ASP, the search task description is given in the form of a request for information (e.g. "Find information about living in India."). Under the task description, two search tiles are presented where each tile shows a separate set of search results for the query. Then the subjects made their selection using a "submit" button.

The subjects (assessors) could choose to perform as many tasks as they wished. To avoid learning effects, we ensured that each assessor was not shown the same task more than once. All studies were performed via a crowd-sourcing platform, Amazon Mechanical Turk.[2] The methods employed to collected the data via this platform is described in Section 3.1.4. The result sets shown on each ASP were pre-crawled offline. To lower assessment burden, subjects were unable to browse outside the ASP, i.e. clicking any links within the result page did not redirect them to external web pages. The snippets on the ASP were the sole source of evidence to assess the SERP quality.

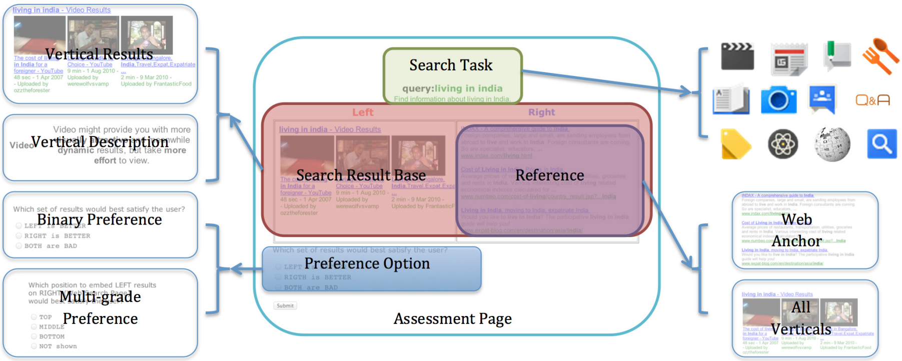


Figure 3.2: Various Components for Manipulations on Assessment Page of Studies 1 to 3 in Understanding Vertical Relevance Assessments

[2] https://www.mturk.com

Table 3.2: Verticals Used in Assessing and Understanding Vertical Relevance.

Vertical	Vertical Engines	Document	Type
Image	Google Image	online images	media
Video	Google Video	online videos	
Recipe	Google Recipe	recipe page	genre
News	Google News	news articles	
Books	Google Books	book review page	
Blog	Yahoo! Blog	blog articles	
Answer	Google Q&A	answers to questions	
Shopping	Google Shopping	product shopping page	
Discussion	Google Forums	discussion thread from forums	
Scholar	Google Scholar	research technical report	
Wiki	wiki.com	encyclopedic entries	
General web	vertical-filtered google.com	standard web pages	

Verticals and Search Tasks

In web search, a vertical is associated with content dedicated to either a topic (e.g. "finance"), a media type (e.g. "images") or a genre (e.g. "news")[3]. In this chapter, we are mainly concerned with the latter two types, which is less well-studied than the former. We use a number of verticals (listed in Table 3.2). Those verticals reflect a representative set of vertical engines used in current commercial aggregated web search engines. Instead of constructing verticals from scratch, we use a representative state-of-the-art vertical search engine for each vertical, as listed in Table 3.2.

Search tasks were chosen to have a varying number and type of relevant verticals. From a preliminary study [Zhou et al., 2011, 2012a] (also described in Chapter 4), we collected annotations of users' preferred verticals for 320 search tasks (from the TREC million query and web tracks, originally derived from search engine logs). The preferred verticals reflect the perceived usefulness of a vertical from the user need perspective, without regard to the quality of the vertical results. This is achieved by instructing assessors to make pairwise preference assessments, comparing each vertical in turn to the reference '*general web*' vertical without viewing any vertical results (including the general web). When making assessments, only vertical names/labels were shown and at least four assessors judged each search task.

[3] A topic-focused vertical may contain documents of various types, standard web pages, images, reviews, etc.

Table 3.3: Example Search Tasks in User Studies for Assessing and Understanding Vertical Relevance.

Search Task Description	Preferred Vertical	Query
I am looking for information on the Welch corgi dog.	Image	welch corgi
Find beginners instructions to sewing, both by hand and by machine.	Video	sewing instructions
I am looking for cooking suggestions of turkey leftover.	Recipe	turkey leftover
Find music, tour dates, and information about the musician Neil Young.	News	neil young
Find information on the history of music.	Books	who invented music
Find information about living in India.	Blog	living in india
Find information on how I can lower my heart rate.	Answer	lower heart rate
I am looking for sources for parts for cars, preferably used.	Shopping	used car parts
Find ”reasonable” dieting advice, that is not fads or medications but reasonable methods for weight loss.	Discussion	dieting
Find information on obsessive-compulsive disorder.	Scholar	ocd
Find information about the Sun, the star in our Solar System.	Wiki	the sun
Find the homepage of Raffles Hotel in Singapore.	General-web Only	raffles

We then select 44 tasks from those 320 search tasks. The selection is to ensure a wide coverage of information needs with different preferred verticals, including those with no preferred verticals. For each of the 11 verticals, we select 3 search tasks where more than 75% of the assessors preferred the vertical. We also select 11 search tasks where none of the verticals were preferred. For each task description, to avoid any bias, we ensured that it did not contain any vertical-explicit request (e.g. "find images for yoga poses."). Twelve representative example tasks (one per preferred vertical) are shown in Table 3.3. Although the search task set is not large, it is sufficient to investigate certain aspects of vertical relevance, upon which large-scale user studies can subsequently be carried out.

Assessment Manipulation

To answer our research questions, each ASP has five components that can be manipulated:

- *search task*: the information need (or search task) that assessors encounter;
- *vertical of interest*: the vertical that is presented for assessments;
- *search result base*: the default type of information presented on the SERP for each ASP;
- *assessment reference*: the reference SERP (one of the two result sets on an ASP) against which an assessor will make a preference;
- *preference option level*: the number of options allowed for an assessment (binary or graded) of an ASP.

Search tasks are manipulated to provide a more complete evaluation of AS information needs. *Verticals of interest* are manipulated to provide a comprehensive evaluation of various verticals for AS. *Search result base* refers to the default type of information provided for assessments and in our study was manipulated for two possible options: search engine description or retrieved search results. Those two options reflect on different influencing factors for assessments. The former type reflects on assessors' pre-retrieval user need perspective (orientation) whereas the latter reflects on assessors' post-retrieval user utility perspective (a combination of orientation, result quality and aesthetic). This relates to **RQ 1** and a detailed design of this manipulation is described in Study 1. *Assessment reference* deals with which information is used as a reference to make the pair-wise preference assessments for a vertical. It is manipulated to investigate whether there is a dependency between (relevant) verticals. We manipulate this to compare *anchor-based approach* and *inter-dependent approach* for **RQ 2 and RQ 3** and a detailed design of this can be found in Study 2. Manipulation of the *Preference option level* provides different levels of granularity for assessors to specify their preference based on the quality of two SERPs. A more fine-grained option (multi-graded) provides more details than other simple options (binary). This is manipulated to investigate how much information is lost when assessors are provided with simpler options. This variable relates to **RQ 4** and its investigation forms Study 3.

We have five independent variables that can be manipulated within an ASP. However, due to a limited budget, instead of using a full factorial design with all the independent variables, we control four variables when investigating one factor. We set the four variables to their most common setting, in a typical AS scenario, and study the change in the behaviour of our assessors when the test variable (which we are currently testing) changes. Except for *search task* and *vertical of interest*, the three other independent variables in our study represent the RQs that we wish to answer:

- *search result base*: pre-retrieval user need (by showing only vertical descriptions) or post-retrieval user utility (by showing retrieved vertical results).
- *assessment reference*: 'general web' anchor (showing only 'general web') or all verticals (including both 'general web' and all other verticals).
- *preference option*: binary or multi-graded.

To measure the effect of the independent variables on users' vertical relevance assessments, we investigate two dependent variables: the **inter-assessor agreement** (measured by Fleiss' Kappa K_F [Fleiss, 1971]) and the **vertical relevance correlation** (measured by Spearman correlation). *The inter-assessor agreement* focuses on measuring the ambiguity (or difficulty) of the vertical relevance assessments. This can give us insights on whether it is difficult for assessors to draw agreement on assessing vertical relevance. *The vertical relevance correlation* measures for two assessment processes, whether one agrees with the other for the search task. This can give us insights on comparing different assessment processes and determining which component of the assessment should be controlled more strictly so that it leads to stronger correlations. We report the results of these two dependent variables for all of our studies.

As we are mainly interested in measuring assessor agreement over assessed preference pairs, instead of employing metrics (e.g. overlap measures [Voorhees, 2000]) to measure inter-assessor agreement on absolute assessments (query-document topical relevance assessment), we used Kappa measure, as prevalently used in previous work [Arguello et al., 2011b]. We select Fleiss' Kappa (denoted K_F) to measures the (chance-corrected) inter-assessor agreement between any pair of assessors over a set of triplets. This allows us to ignore the identity of the assessor-pair because it is designed to measure agreement over instances labelled by different (even disjoint) sets of assessors. Specifically, when M_t is available, for all the assessments for a particular assessment a_j or a set of assessments (A_j) for all assessors U, we can calculate the Fleiss' Kappa over all pairs. Therefore, after calculating K_F for both assessment processes, we can compare their assessment agreement, to obtain insights into assessment difficulty and diversity.

We used Spearman's Correlation as our main tool for our data analysis as it is widely used in IR and it is a powerful statistical method to determine the dependency between two variables of interest (two assessment processes in our work). Due to space limitation, more in-depth analysis of the data (e.g. close manual examination) is left for future work.

Crowd-sourcing Data Collection

Our preference assessment data is collected over the Amazon Mechanical Turk crowdsourcing platform, where each worker was compensated \$0.01 for each assessment made. For each ASP, we collect four assessment points. Running user studies on Mechanical Turk requires quality control and we used two approaches for achieving this: "trap"

HITs and “trap” search tasks. Both these types of trap are only used to identify careless and/or malicious assessors. Following [Sanderson et al., 2010], “trap” HITs are created following a set procedure. Each “trap” HIT consists of a triplet (q, i, j), where either page i or j are taken from a query other than q. We interpreted an assessor preferring the set of extraneous results as evidence of careless assessment. “Trap” search tasks are defined as the search task that contains an explicit reference to a preferred vertical (e.g. “Find information from preferred shopping search results on football tickets”). An assessor who failed to provide preference to an explicitly specified preferred vertical a predefined number of times was treated as assessor. Careless assessors were filtered out and all their assessments were discarded. The actual assessments from the traps were also not used in our analysis.

It is objectively difficult to judge whether one assessor is careless since different users might have different vertical preferences for the same search task, or to estimate the cost associated with different types of errors (e.g. irrelevant verticals, relevant verticals presented at the bottom of the page or bad retrieved results of relevant verticals), as demonstrated by previous work [Arguello et al., 2011b, Zhou et al., 2012c]. As we have two different “trap” approaches and a large percentage of assessments are “traps”[4], we believe that our methodology was able to filter out large percentage of careless assessors.

3.2.2 Study 1: Comparing User Perspective

Study 1 aims to investigate whether vertical relevance derived from different user perspectives correlate with each other. We controlled the *search reference* to ‘general-web’ anchor and *preference option* to binary. Therefore, we provide a vertical of interest and ‘*general web*’ together on an ASP and ask the assessor to provide a binary preference (“left is better” and “right is better”). To avoid over-burdening assessors, we also include an option (“both are bad”) that captures the scenario where a user is confused due to, for example, poor quality SERPs.

For the remaining three independent variables *search task*, *vertical of interest* and *search result base*, we used a full factorial design. We used a total of 44 experimental search tasks that vary in number of preferred verticals, as shown in the upper right in Figure 3.2. Eleven verticals of interest are used. As specified above, the *search result base* variable manipulated the base information for assessments and had two values: “vertical description” and “vertical results”. As shown on the upper left in Figure 3.2, for “vertical results”, the top three items of the vertical search results are returned by the commercial vertical search engine employed. When making assessments, “vertical results” reflects the post-retrieval user utility for each vertical of interest. The “vertical description” did not vary across search tasks. We provided a general description of each vertical that specified the item types provided by the vertical and its unique characteristics (e.g. video results might provide more **visually attractive** and **dynamic** results, but may take **more effort** to view). We aimed to provide an objective description of the typical contents of the vertical to avoid any bias. The vertical relevance assessments derived from “vertical description” reflects a pre-retrieval user need perspective (before retrieving from any verticals, which type of information may satisfy the user needs?).

[4] For example, Study 1 (Section 3.2) contains 18.4% “traps” out of all assessments, which means that approximately for every six assessments made, the assessor encountered one “trap”.

Study 1 had 968 unique conditions (44 search tasks × 2 search result base × 11 verticals of interest). To ensure the quality of assessments, we manipulated 5 "trap" tasks (randomly selected from 11 "trap" tasks, one per vertical) and 1 "trap" HITs for every search task under each search result base. We collected four data points for each condition and in total we had 3872 assessments (4744 assessments including all "trap" tasks and HITs).

3.2.3 Study 2: Effects of Context

Study 2 aims to investigate the impact of the context of other verticals to the relevance assessments of a chosen vertical. Study 2 controlled the *preference option* to binary and *search result base* to "vertical results". For the remaining three independent variables *search task*, *vertical of interest* and *search reference*, Study 2 used a full factorial design. The *search reference* had two possible values: "general-web anchor" and "all-verticals", as shown in the lower right of Figure 3.2. The former used each vertical of interest with 'general web' anchor to form 11 assessment pairs for each search task. The latter used a full possible space of each vertical of interest and all other verticals (including three '*general web*' result sets: top-three, top-four-to-six, top-seven-to-ten) to form a total of 91 assessment pairs for each search task. The assessment pairs of the former is a subset of the latter.

Study 2 had 4004 unique conditions (44 search tasks × 91 assessment pairs). We used the same quality control strategy as for study 1. In total we had 16016 assessments (19620 assessments including all "trap" tasks and HITs).

3.2.4 Study 3: Multi-graded Preference

Study 3 aims to investigate whether it is possible to derive multi-graded preferences using binary preference from a number of users. Study 3 controlled the *search result base* to "vertical results", *vertical reference* to "general-web anchor". We use all of the top-ten '*general web*' results as an anchor in this study. This is to be consistent with the multi-graded assessments we aim to investigate as described below. For the remaining three independent variables *search task*, *vertical of interest* and *preference option*, study 2 used a full factorial design. Specifically, the *preference option* is manipulated to be either *binary* or *multi-graded*, as shown in the lower left in Figure 3.2. Note that this is to compare with the '*general web*' results. For the former, assessors were asked for binary assessments (binary preference, i.e. ToP or NS), while for the latter assessors were asked for multi-graded assessments (ToP, MoP, BoP or NS).

Study 2 had 968 unique conditions (44 search tasks × 2 preference options × 11 verticals) using the same quality control strategy as for study 1. We obtained 3872 assessments (4744 assessments including "trap" tasks and HITs).

3.3 Experimental Results

Our goal is to investigate the correlation of vertical *relevance* when derived from studies with different underlying assumptions. We measure the correlation between two sets of relevance assessments using Spearman's correlation. In each case, we outline whether this correlation is significant.[5] We denote the significance by ▲ (with $p < 0.05$).

Table 3.4: (Study 1) Vertical Relevance using Spearman Correlation with respect to Post-retrieval Approach on a Variety of Influencing Factors (Orientation, Topical Relevance).

Verticals	Image	Video	Recipe	News	Books	Blog
Orientation	0.547▲	0.654▲	0.864▲	0.524▲	0.516▲	0.385▲
Topical Relevance	0.092	0.205▲	0.637▲	0.301▲	0.187▲	0.429▲
Verticals	Answer	Shopping	Discuss	Scholar	Wiki	Average
Orientation	0.563▲	0.610▲	0.305▲	0.450▲	0.404▲	0.529
Topical Relevance	0.354▲	0.264▲	0.571▲	0.393▲	0.484▲	0.356

3.3.1 Study 1

We report the results that compare user vertical relevance I_t from different perspectives. Specifically, whether **(1) orientation (pre-retrieval vertical preference)** and the **(2) topical relevance of post-retrieval search results** affect a user's perception of a vertical relevance. For (2), following a standard TREC-style evaluation methodology, we collected graded topical relevance assessments (highly, marginally and not relevant) for the top search results returned from the verticals (including '*general web*'). Then for each assessment pair (v_i, w), we use $nDCG(v_i) - nDCG(w)$ to quantify the weighted preference of v_i over w based on topical relevance.

We examined the user agreement when assessing the pair-wise preference in both a pre-retrieval and post-retrieval scenario. The Fleiss' Kappa (K_F) obtained for both pre-retrieval and post-retrieval are 0.47 and 0.40, respectively. In both scenarios, the inter-assessor agreement is not high (moderate). This indicates the difficulty (or ambiguity) of AS in general; different users tend to make different decisions regarding the *relevance* of a vertical. A low K_F on a particular query indicates that it is a particularly ambiguous query. Unexpectedly, we observed that there is even more disagreement between assessors when they are allowed to view the results retrieved from each vertical (on each SERP) (post-retrieval setting). In that setting, given that the assessors have more information to make their assessments, one would expect more agreement. However, this is not the case. A number of reasons may cause this. Firstly, it should be noted that as we have only four assessors, the difference in inter-assessor agreement can be substantially affected by one assessor. Secondly, and more importantly, it is possible that providing the search results to each assessor increases the difficulty and ambiguity of the assessment process. This may be due to the fact that the user now has to take more factors into account when making an assessment (pre-retrieval vertical orientation, item relevance, visual attractiveness). These factors may lead to more noisy assessments as each assessor may place different emphasis on these factors. We also calculated the Spearman

[5] We determine the significance by using a permutation test.

correlation of the inter-assessor agreement between the pre-retrieval and post-retrieval assessments. We found that this correlation is high (0.749), indicating that in both scenarios (pre-retrieval and post-retrieval) the assessors encounter difficulty with the same queries.

Furthermore, we report the Spearman correlation of the two influencing factors (orientation and topical relevance of items) with respect to the post-retrieval vertical relevance for a variety of verticals. The higher the correlation, the more important the factor is in influencing the utility of the search results (from the user point of view). This is shown in Table 3.4. We can observe that the average Spearman correlation of orientation (pre-retrieval) and topical relevance with respect to post-retrieval vertical relevance over all verticals is 0.529 (moderate) and 0.356 (low), respectively. These correlations are not particularly high (but all are significant) for both influencing factors. Generally, *orientation* is more highly correlated with the utility of a set of search results than *topical relevance*. This demonstrates that neither factor can solely determine the user's perception of the utility of the search results. In addition, in our data, the type of vertical (orientation) is more important for the search result utility than the topical relevance of the search results.[6] When we analyze the *orientation* of each vertical, we observe that some of the verticals obtain comparatively high correlation ('*Video*', '*Recipe*' and '*Shopping*') whereas others obtain comparatively low correlation ('*Blog* and '*Discussion*). This suggests that some verticals are inherently more ambiguous in terms of their usefulness for the search task than others.

For *topical relevance*, we observe that the topical relevance of retrieved results for the '*Image*' vertical does not contribute significantly to the search results utility. An in-depth examination showed that this can be explained by the lack of variability of the topical relevance. We observe that most returned image results are topically relevant. Conversely, the topical relevance of the items of other verticals ('*Blog*', '*Discussion*') contributes a larger degree to the utility of a SERP. This is because for those verticals, the results are too similar to '*general web*' results and in this case, topical relevance is the most important aspect for search utility (as in traditional web search). For '*Recipe*', topical relevance correlates highly both with orientation and search utility. This is because '*Recipe*' is more likely to contain relevant results only when user are oriented to that vertical.

Thirdly, as we are more concerned with highly relevant verticals, we investigate whether the top relevant verticals are the same for pre- and post-retrieval scenarios. We extract the top-three most preferred verticals from both assessment scenarios and compare them. We calculate the overlap between them and the results are shown in Table 3.5. There is generally some overlap between vertical relevance for around 90% of the queries. In addition, in 56.8% of the search tasks at least two out of three relevant verticals are in common, when relevance is derived from the different assessment methods (pre- and post-retrieval assessments).

Finally, we investigate whether there is an aesthetic bias for verticals that present more visually salient results ('*Image*', *Video*' and *Shopping*' in our study). We compare the number of occurrences of those verticals that appear within the top-three verticals for

[6] Note that due to our selection of vertical search engines (highly performing verticals) where most vertical search results contain topically relevant items for most of the search tasks, our results are biased to this scenario and might not generalize when vertical search engines perform badly.

Table 3.5: Overlap of the Top-three Relevant Verticals for Pre-retrieval (Orientation) and Post-retrieval (Search Uutility) for the same Search Tasks.

Overlap	3	2	1	0
Num of Tasks	5	20	14	5
Fraction	11.4%	45.4%	31.8%	11.4%

various search tasks. Consistent with previous work, we found there is an aesthetic bias in user's perception of the utility of the search results. There are in total 21 occurrences of those verticals appearing within the top-three verticals for all search tasks in the post-retrieval case, compared with 11 occurrences within the pre-retrieval case.

To summarize, Study 1 shows that both *orientation* and *topical relevance* contribute significantly to the search result utility, whereas the impact of *orientation* is more important. In addition, there is an aesthetic bias to user's perception of the search results utility.

3.3.2 Study 2

In Study 2, we manipulated the assessment reference for each vertical of interest. Again, the reference is manipulated by presenting only general-web anchor results (anchor-based approach) in one approach and all vertical results (inter-dependent approach) in a separate approach. To derive I_t using assessments M_t obtained for each search task, we used an existing approach. For the *anchor-based approach*, we ranked all the verticals of interest based on the percentage of assessors' preference over 'general web' anchor. For the *inter-dependent approach*, we used Schulze voting method [Arguello et al., 2011b]. We report the results comparing user's vertical relevance I_t from both the anchor-based approach and the inter-dependent approach. For the former, we vary the quality of the 'general web' anchor by using different result sets (Web-1: top 1-3 items, Web-2: top4-6 items or Web-3: top7-10 items). We aim to investigate whether there are significant differences between them.

We look at the user assessment agreement. The Fleiss' Kappa (K_F) obtained for the anchor-based and inter-dependent approaches are 0.40 and 0.42, respectively. The user assessment agreement is not high (moderate) and, generally, there is not much difference between the assessment agreement of the two approaches. The slight increase of user agreement for assessments in the inter-dependent approach might be due to the comparative ease in assessing some vertical-pairs, over assessing vertical-anchor-pairs.

We show the query-specific Spearman correlation of the anchor-approach using different anchors (Web-1, Web-2, Web-3) with respect to the inter-dependent approach. The results are shown in Table 3.6. We can observe several important trends. Firstly, the correlation between the anchor-based and inter-dependent approaches is *moderate*. From closer examination, we see many "exchange" between verticals of similar intended level and most of these "exchanges" occur within lowly vertical relevance level. As we are more concerned with highly intended verticals, similarly to Study 1, we report the overlapped top relevant vertical between the two approaches in Table 3.7. Generally the overlap of the top-three relevant verticals between these two approaches is quite high (more than 70% of the search tasks have the same perception of at least two out of three relevant verticals).

Table 3.6: (Study 2) Spearman Correlation of Vertical Relevance Derived between Anchor-based Approach (using anchors Web-1, Web-2, Web-3) and Inter-dependent Approach.

Anchor	Web-1	Web-2	Web-3	Average
Correlation	0.626▲	0.515▲	0.579▲	0.573

We observe that although there are differences between the approaches that use different anchors, the differences are not large in general (all moderate correlations). Web-1 generally correlates higher than Web-2 and Web-3, and there is not much differences between Web-2 and Web-3. This is quite surprising. We assumed that the change of topical relevance level of the anchor results[7] would result in a change of a user's perception of the results utility. However as this is not the case, we suspect that this can be explained by the finding in Study 1, where when presented with a 'general web' anchor, it is the *type* of information that leads to a more significant impact on the quality of the result set, indeed more so than topical relevance.

Finally, to demonstrate the interaction between verticals, an analysis of the difference between the inter-dependent ranking and anchor-based (Web-1) ranking suggests that context matters, i.e. the relevance of the latter vertical diminishes when the former vertical (context) is shown in advance. We analyse this by finding the most frequent discordant pairs of verticals (v_i, v_j) within the two approaches. All the candidate pairs consist of verticals of interest occurring within the top verticals for at least one approach. We found that most pairs are concordant with each other but there are about 14% of discordant pairs. Specifically, there are several distinct discordant pairs that consistently occur for different number of top results (3 to 6). These pairs are ('*Answer*', '*Wiki*'), ('*Books*', '*Scholar*'), ('*Answer*', '*Scholar*'). For example, ('*Answer*', '*Wiki*') pair means that when '*Answer*' is presented before '*Wiki*', the relevance of '*Wiki*' is diminished. This might be explained by the fact that once a direct answer is available, reading a long wiki article will provide less utility to the user. These results demonstrate that the context of other verticals can diminish the utility of a vertical. This finding requires further examination.

3.3.3 Study 3

We investigate how various thresholding approaches can be used to accurately derive multi-graded vertical relevance for the anchor-based approaches. We also apply this to the Schulze voting method for the inter-dependent approach [Arguello et al., 2011b].

[7] We found that the averaged nDCG values satisfy $nDCG(Web\text{-}1) > nDCG(Web\text{-}2) > nDCG(Web\text{-}3)$ based on topical relevance.

Table 3.7: Overlap of the Top-three Relevant Vertical for the Anchor-based Approach (Web-1) and the Inter-dependent Approach on same Search Tasks.

Overlap	3	2	1	0
Num of Tasks	12	19	10	3
Fraction	27.3%	43.2%	22.7%	7.8%

For each search task, based on the multi-graded assessments for each vertical v_i (assessed by four independent assessors), we first derive the ground-truth of the "perfect" embedding position[8] (and corresponding "perfect" page). To achieve this, we assume that there is a continuous range for each grade ($[3, 4]$ for ToP, $[2, 3)$ for MoP, $[1, 2)$ for BoP and $[0, 1)$ for NS). We assign each grade the medium of its corresponding range as its weight (3.5 for ToP, 2.5 for MoP, 1.5 for BoP and 0.5 for NS). Then for four assessors' judged grade, we decide the "perfect" position by calculating the expected assessed grade's weight and finding its corresponding fitted grade range.[9]

For the anchor-based approaches, we use a set of thresholding settings (for binary assessment, this is the fraction of assessors that deem the vertical as relevant) for ToP, MoP, BoP, respectively. For a given vertical, when the fraction of its assessors' assigned "relevant" is larger or equal to the weight assigned for a given grade, we treat that vertical as that specific grade. We vary those thresholding settings for different risk-levels: risk-seeking (0.5, 0.25, 0), risk-medium (0.75, 0.5, 0.25) and risk-averse (1, 0.75, 0.5). As described above, we also use another existing approach (Schulze voting method [Arguello et al., 2011b]) for the inter-dependent approach.

Firstly, we look at the user assessment agreement. The Fleiss' Kappa (K_F) obtained for binary and multi-graded approaches are 0.40 and 0.35, respectively. The agreement of multi-graded assessments is not high.[10] From a closer examination, we found that this might result from each assessors' unique preference of verticals and their risk-level [Zhou et al., 2012b] (i.e. their willingness to take risk to view more irrelevant verticals). Some of the assessors tend to choose more verticals to be shown at earlier ranking (e.g. ToP, BoP) while others are more careful and select verticals to be shown on the SERP only when they have a high degree of confidence.

Secondly, for each approach used to derive vertical relevance from binary assessment, we obtain its corresponding optimal page (with '*general web*' results Web-1, Web-2, Web-3 and verticals that are shown). Then we calculate the Spearman correlation of this page with the ground-truth page derived from the multi-grade assessments. The results are shown in Table 3.8. As we are concerned with how each binary approach can be used to derive accurate multi-graded assessment, we also calculate the precision of each

[8] Note that this "perfectness" of embedding position and page is likely to be sub-optimal. This is because the multi-grade assessment methodology does not capture the context of other verticals.

[9] For example, when two, one, one and zero assessors assign ToP, MoP, BoP and NS, respectively, we obtain the expected weight of grade $(2 \cdot 3.5 + 1 \cdot 2.5 + 1 \cdot 1.5 + 0)/4 = 2.75$) and therefore its "perfect" embedding position is MoP (as $2.75 \in [2, 3)$).

[10] Note that this K_F agreement is not directly comparable to others as the number of assessment grades changes.

Table 3.8: (Study 3) Spearman Correlation of Optimal Pages derived from Binary Assessments and Ground-truth Page derived from Multi-grade Assessments, and Precision (for each grade ToP, MoP and BoP).

Binary Approach	risk-seeking	risk-medium	risk-averse	Schulze voting
Correlation	0.135	0.411▲	0.292▲	0.539▲
prec(ToP)	0.30	0.52	0.74	0.67
prec(MoP)	0.18	0.31	0.43	0.25
prec(BoP)	0.09	0.26	0.37	0.39

binary approach with respect to the multi-grade ground-truth.

We notice several important trends. Firstly, most of the binary approaches (risk-medium, risk-averse and Schulze voting) are all significantly correlated with the multi-graded ground-truth. However, the correlations are mostly moderate. It is not surprising that Schulze voting method performs the best, as it uses more assessments (91 assessments) compared with other binary approaches (11 assessments) as well as being more robust to noise. It is also interesting to observe that the risk-medium approach performed second best, which is consistent with our observation that different assessors have different risk-levels. An extreme approach (risk-seeking or risk-averse) is more likely to satisfy only a small subset of assessors while frustrating others. Secondly, when focusing on the precision of each approach for each grade (ToP, MoP and BoP), we can observe that generally, risk-averse performs best, followed by Schulze voting, risk-medium and risk-seeking approaches. This is because the risk-averse approach is more careful when selecting verticals; it only selects verticals (as relevant) when highly confident (large fraction of user's preferences) of this.

3.4 Conclusions and Discussions

Our objective of this chapter is to investigate whether different underlying assumptions made for vertical relevance affects a user's perception of the relevance of verticals. Our results indicate that relevant verticals derived from different assumptions do correlate with each other. However, the correlation is not high (either moderate or low in many cases) as each assumption focuses on different aspects of vertical relevance. With respect to RQ1, both *orientation* (pre-retrieval user need) and *topical relevance* (post-retrieval topical relevance) correlates significantly with the post-retrieval search results utility. The impact of orientation is comparatively more significant (moderate) than topical relevance (low). In addition, there is an aesthetic bias to a user's perception of search results utility. With respect to RQ2, we conclude that the context of other verticals has significant impact on the relevance of a vertical. With respect to RQ3, we found that it is possible to employ a number of binary assessments to predict multi-grade assessments and the correlation of the derived optimal pages is significant (moderate). Using a larger number of assessments (e.g. Schulze voting) contributes to more accurate estimation of multi-grade assessments.

Our results have important implications for aggregated search and in general, evaluation in IR. The moderate correlation between different vertical assessments indicates the need to re-evaluate previous work on vertical selection, based on the assessments (and corresponding assumptions) used. The conclusion drawn from one type of assessments (e.g. VS approach A performed better than B) might not hold for another type of assessments. Researchers need to be careful when drawing conclusions regarding vertical relevance.

Our results have implications for work in vertical selection. As discovered in Study 1, *orientation* has a larger impact on user's perception of the search results utility than topical relevance, which implies that vertical evidence derived from the user need perspective (e.g. query logs) might be more effective at predicting a user's relevant verticals than collection-based estimation (e.g. traditional resource selection methods). In addition, Study 1 implies that for some verticals (e.g. *Video*', *Recipe*' and '*Shopping*'), the VS system generally would have more confidence in returning them as relevant (due to their *orientation*). On the contrary, the VS system should be more careful when returning other verticals (e.g. '*Blog*' and '*Discussion*' results). We are not saying that some verticals ('*Video*') are more useful than others ('*Blog*' and '*Discussion*'); we note that it is easier to *predict* the usefulness of some verticals for an "average" query.

Our results have implications with respect to procuring assessments for aggregated search. In Study 2, we showed that fewer binary assessments (anchor-based approach) correlate moderately with more binary assessments (inter-dependent approach). In Study 3, we showed that moderately correlated multi-graded relevance assessments can be obtained by using a number of binary assessments. As different assessment methodologies involve differing amounts of effort (number of assessments, information load when assessing), there is a need for analyzing both the utility and effort involved in different assessment methodologies so that assessments can be obtained in a more efficient way. In addition, by exploring verticals on aggregated search pages, binary preference of vertical over web results can be obtained/derived by mining query logs [Ponnuswami et al., 2011b].

Our work also has the following limitations. Firstly, although we have shown that topical relevance has significant impact on user's perception of search results utility, we have not explored how this impact changes according to the different levels of topical relevance, and how it interacts with orientation. Similarly, a comprehensive analysis on aesthetic bias is also needed. Secondly, at the moment we assume a blended presentation strategy, i.e. interleaving vertical results into the web results (ToP, MoP, BoP and NS). Other ways of combining results are possible, for example showing blocks of results on the right side of the page. Finally, the assessments have been obtained by showing only vertical search result snippets to the users, without presenting the actual information items. As this assessment is depended on only the snippet quality, we should examined the impact of this further.

This chapter focused primarily on analyzing how to gather vertical-level relevance assessments. This is crucially important and helps us better understand the relevance from the perspective of vertical. We could be able to evaluate key components such as vertical selection and result presentation using this vertical relevance assessments. However, in order to further understand and evaluate item selection, we require further assessments on the documents within the verticals. Therefore, we further follow-up on the work presented in this chapter by building a TREC-style aggregated search test collection that collect document-level relevance assessments.

4

Building a Test Collection by Reusing

In the previous chapter we focused on collecting vertical relevance assessments. In this chapter, we start our investigations by studying how to efficiently and cheaply build a test collection for aggregated search. A test collection for aggregated search requires a number of verticals, each populated by items (e.g. documents, images, etc) of that vertical type, a set of topics expressing information needs relating to one or more verticals, and relevance assessments, indicating the relevance of the items and their associated verticals to each of the topics. Building a large-scale test collection for aggregate search is costly in terms of time and resources.

This chapter addresses our research questions **RQ 5** to **RQ 6**, as specified in Section 1.2.1. We propose a methodology to build such a test collection reusing existing test collections. With this created test collection, we address the following research questions:

> **RQ 5**: Can we reuse existing test collections to construct a test collection for aggregated search?
>
> **RQ 6**: Is the constructed test collection reliable? What is the impact of misclassification (of items into verticals) to the evaluation of systems?

Following this methodology of "reusing", a test collection is built, which allows the investigation of aggregated search approaches and evaluation in a timely fashion and with the required focus. In this way, as new, more focused verticals become available, they can be seamlessly integrated into the existing collection. It should be noted that our focus is not to replace the test collection creation methodology used in TREC, but rather utilise a similar methodology to create a practically useful, reliable, and consistent test collection for the aggregated search community. We also report on experiments that show that the methods used to build our collection lead to a reliable and reusable collection.

The remainder of this chapter is organized as follows: Section 4.1 provides some background. Section 4.2 describes our "reusing" methodology to construct a test collection for aggregated search from existing test collections. Section 4.3 describes the stages and design decisions involved in using a SVM classifier to classify documents (items) into various types. Section 4.4 details experiments carried out to investigate the consistency of the constructed test collection and discuss its reliability by experimenting with twelve aggregated search systems. Finally, Section 4.5 outlines our conclusions.

4.1 Related Work

Current test collections have become extremely large (e.g. Clueweb09) to reflect the much larger amount of information in many of today's retrieval scenarios. As a result, the idea of reusing test collections has been proposed. Some researchers [Clarke et al., 2008] have reused an existing Q&A test collection to generate a test collection to investigate diversity in IR. Others [Carterette et al., 2010] have developed means to quantify the reusability of a test collection for evaluating a different retrieval scenario than that originally built for.

The most time-consuming part of generating a test collection is the creation of relevance judgments. Although methods to alleviate this problem have been proposed (e.g. formally selecting a subset of the most promising items to be judged [Carterette et al., 2006] or using crowd-sourcing techniques [Arguello et al., 2011b]), judging a set of items (often documents), and in particular, heterogeneous documents from a variety of sources, remains an extremely tedious task.

Other related work of creating a test collection for aggregated search includes collecting pair-wise judgments on information originating from different verticals (vertical blocks) via crowd-sourcing [Arguello et al., 2011b], or inferring judgments from query-log [Ponnuswami et al., 2011b]). Our approach emphasizes the *reuse* of collections and judgments, and furthermore, leads to a reusable collection. In homogeneous federated search, test collections have been developed reusing existing test collections by partitioning different text-based corpora into a number of sub-collections. The partitions [Shokouhi and Si, 2011] are generally based on topicality, publication source, date, or domain. In desktop search, test collections [Kim and Croft, 2009] have been created by collecting different types of information (e.g. email, web-page, office documents, etc.) for individuals.

Importantly, yet lacking in the aggregated search domain in the early stage of this thesis, the construction of a test collection allows different components of an aggregated search system to be systematically evaluated on a stable collection, that will be valuable for the research community. In particular, our methodology is based on reusing existing web collection (ClueWeb09[1]) and multimedia collections, and the vertical partitioning reflects a realistic scenario (i.e. realistic verticals) on the web. The items contained in the verticals are of different media (e.g. image and text) and genres (e.g. Blog, News, Wiki, etc.). In addition, all the topics in our test collections simulate real information needs as they come from search engine query-logs. These, although not perfect, more accurately reflect an aggregated search scenario. It should be noted that at this stage we do not investigate the temporal nature of verticals in this chapter (and this thesis).

4.2 Methodology

In this section, we describe the methodology of *reuse* adopted herein to construct an aggregated search test collection. The methodology can be categorized into several steps:

1. First, we *define* the verticals that we want to investigate (Section 4.2.1).

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

2. Then, we decide which existing test collections to use and how to *simulate* verticals (i.e. by classification) (Section 4.2.2).

3. Thirdly, we *identify* a set of topics, from existing ones, that could be satisfied by documents that are contained in several (one or many) simulated verticals (Section 4.2.3).

4. Furthermore, we discuss how existing *relevance assessments* can be used correctly so that the aggregated collection remains reliable.

4.2.1 Defining Verticals

In web search, a vertical is associated with content dedicated to either a topic (e.g. "finance"), a media type (e.g. "images") or a genre (e.g. "news")[2]. In this chapter, we are mainly concerned with the latter two types, which is less well-studied than the former (e.g. topic-focused distributed collections have been studied in federated search [Si and Callan, 2003a]). Consistent with existing web search engines, we consider the verticals listed in Table 4.1. These verticals can be simulated by existing test collections (mainly web-based and multimedia collections), as we show in Section 4.3. The last vertical in Table 4.1, "general web", consists of the standard web search pages, that form the majority of search results [Murdock and Lalmas, 2008]. It is to these results that results from other verticals are added, if relevant [Arguello et al., 2009].

4.2.2 Simulating Verticals

For the purposes of building our aggregated search collection, two main types of existing test collections are available. The first type of collection are those that could be used in their entirety, to simulate a vertical. The second type of collection are those that need to be decomposed into parts, each of which could be used to simulate a vertical, or part thereof. Examples of the latter include large-scale web collections, comprised of documents that are not only standard web documents, but of various genres (e.g. news, wiki, blogs, etc). Documents in such a collection are more problematic as they need to be classified into a genre, and then added to the corresponding vertical.

4.2.3 Identifying Topics

Now we must identify a subset of the topics (from all available topics) that could reflect concrete search scenarios in aggregated search. Following [Arguello et al., 2009], this subset should consist of approximately 1/4 of the topics for which only the "general web" vertical is of high *vertical intent*, and 3/4 for which more than one vertical (including "general web") is of high *vertical intent*. At this stage, we must clarify the concept of "*vertical intent*" when referring to a vertical. We define two criteria to determine the *vertical intent* of a vertical:

[2] A topic-focused vertical may contain documents of various types, standard web pages, images, reviews, etc.

Table 4.1: Verticals that are Simulated in Building a Test Collection by Reusing

Vertical	Document	Type
Image	online images	media
Video	online videos	
Recipe	recipe page	genre
News	news articles	
Books	book review page	
Blog	blog articles	
Answer	answers to questions	
Shopping	product shopping page	
Discussion	discussion thread from forums	
Scholar	research technical report	
Reference/Wiki	encyclopedic entries	
General web	standard web pages	

1. **Topical relevance**, i.e. the vertical should contain at least one topically relevant document (i.e. it should be capable of satisfying the user's need in a topical manner).

2. **Vertical orientation**, i.e. the degree to which a specific type of information, originating from one specific vertical, satisfies a user's information need (e.g. images are highly oriented to the topic "photographs of flowers").

We state that a topic has a high *vertical intent* to a specific vertical only when both criteria are satisfied. Therefore, to identify a set of usable topics, we must first identify verticals that contain at least one relevant item for a topic. Then, we must identify if those verticals have a high *vertical intent* for each of the queries.

4.2.4 Using Existing Relevance Assessments

Reusing existing relevance assessments is one of the most problematic areas when it comes to creating an aggregated search collection. As topics for one simulated vertical, typically do not overlap with topics from another, it is difficult to collect a large set of topics that span multiple verticals. To avoid a situation whereby whole verticals have not been assessed (for relevance) for particular topics, we have used one large collection of heterogeneous documents (ClueWeb). We do, however, use two different media type collections (e.g. image and video). These are of a different media type and the majority of ClueWeb topics do not have corresponding relevance assessments available within these collection. This is somewhat problematic. However, one way to minimise this impact is to manually judge the vertical intent of each query, and then perform relevance assessments only on collections that might be useful for that query. In this work, this incompleteness is minimised and we assume that the image and video vertical verticals have a very low *vertical intent* for the queries originating from the ClueWeb query set.

Table 4.2: Description of Collections, Topics and Qrels Used For Building a Test Collection for Aggregated Search.

Collection	Type	Num of docs	Track	# Topics
ClueWeb09(B)	general web	50,220,423	TREC Web, Million Query	785
ImageCLEF	image	670,439	ImageCLEF, WikiMM	223
TRECVID	video	1,253[3]	TRECVid	268
Total		50,892,115		1,276

4.3 A Test Collection for Aggregated Search

In this section, we describe the actual construction of our test collection. We will describe our document classification approach (Section 4.1), topic identification method (Section 4.2) and statistics of the created test collection (Section 4.3), respectively.

4.3.1 Document Classification

Table 4.2 lists the collections and topics used in our aggregated collection. The majority of our documents come from the ClueWeb collection and, therefore, need to be classified into a specific vertical. We now describe the classification approach used for this web-based collection. Genre classification is not new in the community [Santini, 2007] and our aim is to demonstrate the feasibility of the approach (rather than thoroughly investigating how to improve genre classification). Our classification can be categorized into two steps:

1. Classifying the unlabeled documents using a *machine learning genre classifier*.
2. Increasing the accuracy of the classification of documents from Step 1 using existing vertical search engines.

For Step 1 (machine learning classification), we use a two-stage classifier. The first stage filters out pages that do not occur within the domains that are known to be associated with a vertical (e.g. www.recipies.com for the recipies vertical), and the second step uses a SVM classifier with other features of the page. In the first stage, we filter out the low-quality websites/domains for each vertical, since in our preliminary experiments those “poor quality” websites have been empirically shown to contribute a lot to the misclassification. We constructed this filter by using a website ranking service $Alexa$[4]. For each vertical, we manually find the top 100 ranked domains (e.g. www.recipies.com) that exist in our collection, and only web pages from those domains are candidate documents to be classified.

[3] Each video is a mixture of many events/shots (normally more than one hundred) that can be further segmented.

[4] www.alexa.com

Table 4.3: Genre Classification of Documents into Verticals – Confusion Matrix

Vertical	Recipe	News	Books	Blog	Answer	Shopping	Discuss	Scholar	Wiki	Web
Recipe	**0.74**	0.00	0.02	0.04	0.01	0.00	0.01	0.00	0.00	**0.18**
News	0.00	**0.60**	0.00	**0.05**	0.01	0.01	0.00	0.00	0.00	**0.33**
Books	0.01	0.01	**0.53**	0.04	0.00	**0.10**	0.00	0.00	0.01	**0.30**
Blog	0.02	0.03	**0.06**	**0.57**	0.03	0.02	0.02	0.00	0.01	**0.24**
Answer	0.02	0.02	0.00	0.00	**0.72**	0.00	0.02	0.01	0.00	**0.21**
Shopping	0.02	0.03	**0.07**	**0.07**	0.02	**0.58**	0.04	0.01	0.00	**0.16**
Discuss	0.00	0.02	0.01	0.03	**0.05**	0.02	**0.78**	0.00	0.00	**0.09**
Scholar	0.00	0.00	0.01	0.00	0.00	0.00	0.00	**0.81**	0.04	**0.14**
Wiki	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.00	**0.89**	**0.08**
Web	0.00	0.03	0.01	0.04	0.00	**0.05**	0.02	0.00	0.00	**0.85**

In the second stage, we use a multi-class SVM[5], which is known to perform well for this genre classification task, on these candidate documents. Both textual and structural features are used in the SVM. The textual features include the term-frequencies in various parts of the web document (URL, title, meta-data and full document), genre-based symbols (e.g. the "?" symbol contained in help documents), and named-entity features. Structural features include *html* tag frequency (e.g. list, form, image tag count), links (e.g. number of outlinks). The SVM classifier was trained using five-fold cross validation where the training and testing data consisted of approximatively two hundred manually labeled documents of each genre, and one thousand manually labelled documents of the "general web" genre. Note that all those training web pages exist in the collections we use (i.e. ClueWeb09 B). We have found that approximately 25% of the pages in the collection can be labeled as a non-web vertical by utilizing this two-stage classifier method.

To boost the accuracy of the classification provide in step 1, we submitted the "titles" of all of the classified pages to existing vertical search engines. Therefore, for each document (web page) from the classified set, we submit its "title" to all the corresponding state-of-the-art vertical search engines [6] by using the strict matching retrieval function (i.e. the exact title has to appear in the document.). Then, if the URL of the document (unique identifier) appears in the top 20 results (empirically shown to be sufficient), we relabelled the page with the corresponding vertical. We have found that 18.9% of the classified pages (i.e. those already classified into verticals) were re-classified using this method. This relabelling step (step 2), only affects about 18.9% of the initially labelled documents, but improves the accuracy of the classifier by over 10%.

After those two steps, all the documents in the ClueWeb B collection have been classified as either belonging to a vertical or the default "general web" vertical. Table 4.3 shows the confusion matrix for our genre classification (remembering that these results are generated from 200 manually labelled documents from each vertical using five-fold cross validation). The right-to-left diagonal shows the percentage of correctly classified documents of each type. We achieve an average accuracy of 70.7% (varying from 53% to 89%). Importantly, most mis-classifications are placed into one vertical (i.e. the "general web"). This is not surprising as "general web" is the default genre [Murdock and Lalmas, 2008]. This should not affect our work as documents from the "general web", as is the case for major search engines, form the high majority of search results [Murdock and Lalmas, 2008]. In addition, the overall misclassification remains low and it is comparable to state-of-the-art ([Santini, 2007, Kanaris and Stamatatos, 2009]). Furthermore, this classification reflects the real scenario of vertical creation on the web. Regardless, our experimental section (Section 4.4) will revisit the impact of this classification process.

4.3.2 Identifying Topics

In Section 4.2.3 we defined vertical intent as being related to both *topical-relevance* and *vertical orientation*, and therefore, we must identify a set of topics that are associated to multiple verticals that contain both of these criteria.

[5] http://svmlight.joachims.org/svm_multiclass.html

[6] We use Google News, Blog, Recipe, Shopping, Book, Answer, Discussion, Scholar, and Wiki.com for Reference Search.

Identifying topics associated with multiple topically-relevant verticals

First, we wish to identify topics for which topically relevant documents exist in multiple verticals. This is not problematic for the ClueWeb B collection as we have automatically classified documents into different verticals, and therefore, relevant documents for a topic will be classified into different verticals. However, for the multi-media collections (i.e. image and video), we identify topics that are statements of the same, or a very similar, information need, as those that exist in the ClueWeb topic set. Therefore, we represent each topic as a weighted vector of its *title* terms (i.e. using $tf \cdot idf$) and the cosine similarity is then used to compare topics. Any pair of topics for which the cosine similarity is above a threshold γ are candidate topics. We then manually judged all candidate topics, using the *description* and the *narrative* fields. This yielded two video topics that had a similar information need to those in the ClueWeb B topic set.

Identifying topics with high vertical-orientation

To determine this topic set, first, we make an assumption that highly oriented verticals for a topic should contain above a certain *threshold* of relevant items. Therefore, to define a threshold for each vertical, we analyse a query log (i.e. the AOL log). We identified a set of queries in this log that were highly orientated to a particular vertical (v_i). We identified queries that were highly orientated to a vertical by simply finding queries with an explicit vertical label (e.g. if the vertical name “recipe” appeared in the query “pork chops recipe” we deemed it a recipe query). We also used the main sub-query, created by removing the vertical label (e.g. “pork chops”), as a highly orientated query. Then for each query, we then calculated the fraction of clicks that linked to pages in that vertical (v_i), compared to the number of total clicks for that query. These fractions were then averaged over all queries that were identified as highly *orientated* to a vertical. Given that a click is a noisy estimation of relevance, this fraction gives us an estimation of the number of relevant documents that must be in a vertical before the vertical is deemed highly-orientated. Finally, for our simulated collection, a vertical was deemed *highly-orientated* when it contained over this threshold of relevant documents. Using this process, 270 queries were created that had multiple vertical intents. We compared these vertical intents with those of two human annotators for a subset of queries and found a high degree (60%) of overlap.

4.3.3 Created Test Collection

In this section, we describe the obtained collection, and in addition we answer the following questions. With respect to **RQ 5** on the validity of the “reuse” methodology, we aim to answer: (Q1) Are there enough verticals and are they representative? (Q2) Are there sufficient topics with multiple vertical intents? (Q3) Are the judgments complete and consistent enough?

Table 4.4 shows the document statistics of our aggregated test collection in terms of verticals defined. In total, we have more than 50 million documents. General web documents are prevalent, thus mimicking aggregated search scenarios. A total of twelve verticals are simulated and many are common to the usual 16 verticals[7] found in current search engines (Google, Yahoo and Bing). We have simulated many of the verticals that are prevalent in web search engines and we have simulated more verticals than some search engines. Thus, the number of verticals and their constitution are sufficiently representative to study aggregated search (Q1).

Table 4.4: Document Statistics of the Aggregated Search Collection (Verticals).

Verticals	Recipe	News	Books	Blogs	Answer	Shopping
Ratio	0.3%	3.0%	1.4%	3.8%	0.6%	1.6%
Verticals	Discussion	Scholar	Reference	Image	Video	Web
Ratio	1.1%	0.1%	12.6 %	1.3%	0.0%	74.2%

Statistics relating to the final set of topics and qrels are shown in Table 4.5. In total, 320 topics are available for testing, which is larger than the minimum recommended number of topics in other areas of IR (i.e. 50) [Zobel, 1998]. Also, 69.7% of the topics have two vertical *intents* and 6.2% of topics have three or more vertical intents. These statistics are somewhat comparable to those from [Arguello et al., 2009], obtained from real data. The distribution of topics per vertical is shown in Table 4.6, which also conforms to that of [Arguello et al., 2009]. In summary, the topics forming our test collection are sufficient and reflect a variety of vertical intents (Q2).

Table 4.5: Statistics of Topics and Qrels for the Created Aggregated Search Test Collection.

Statistics	number/ratio
number of topics	320
average relevant docs per topic	26.0
average relevant verticals per topic	1.83
ratio of topics with only “general web” intent	24.1%
ratio of topics with two vertical intents	69.7%
ratio of topics with more than two vertical intents	6.2%

We now discuss the consistency of the relevance judgments. The relevance judgments should be made based on the same criteria for all the documents across all verticals. For all of topics with judgments made on web collection (Clueweb09 B), the consistency is ensured since all the judgments are were made on one collection. However, for the topics with multimedia intent (e.g. topic co-exists in video and web collections), this is not the case. Nonetheless, we can still ensure some level of consistency, because for our chosen topics (Section 4.2.2), we manually judged whether two topics had the same

[7] 16 common verticals include News, Image, Video, Blog, Discussions, Answer, Reference, Maps, Books, Updates, Scholar, Shopping, Financial, Local Listings, Weather, Web.

information need. Thus, the consistency of the relevance judgments have been controlled to a satisfactory degree (Q3).

Table 4.6: Percentage of Topics Assigned to Each Vertical in the Created Aggregated Search Test Collection.

Verticals	recipe	news	books	blogs	answer	shopping
Percentage	3.8%	4.1%	3.8%	5.3%	4.7%	5.6%
Verticals	discuss	scholar	wiki	image	video	web only
Percentage	0.3%	0.0%	54.7%	0.0%	0.6%	24.1%

4.4 Experiments

In this section, we aim to answer **RQ 6** on the reliability of the test collection with respect to the mis-classified documents into the verticals. We used a classifier to assign documents to verticals and, therefore, some documents may be incorrectly assigned to a vertical. We need to assess the impact of this. We now describe an experiment carried out to evaluate the effect that document misclassification has on our newly created test collection. We create different versions of the test collection, where the only difference is that we intentionally mis-classify a certain percentage of the documents. Then, having created these modified collections, we investigate whether the ordering (based on an effectiveness metric) of a number of different aggregated search systems is preserved (or at least correlated) when run on these different versions of the test collection (i.e. collections that have various levels of misclassification). If the ordering of the systems is preserved, or highly correlated, we can conclude that the effect of misclassified documents on our created collection is minimal.

4.4.1 Simulating Misclassified Documents

We create several "misclassified" collections where we reassign the documents into incorrect verticals. For topics that have at least one vertical intent (excluding "general web"), for each specific vertical, we distribute a percentage of its documents uniformly into the remaining incorrect verticals. We iterate this process across all verticals (excluding "general web"). Therefore, according to different misclassification rates" (from 5% to 50%), we create a set of "misclassified" test collections. We also create another test collection (called "random") by randomly assigning documents into verticals. This corresponds to a random classification of documents with regard to the vertical contents.

4.4.2 Simulating Aggregated Search Systems

The key components in aggregated search systems are *vertical selection*, *item selection* and *result presentation*. We generate a set of aggregated search systems by combining different variants of each component. For vertical representation for the *vertical selection*, we use one complete and one incomplete representation that uses query-based sampling [Callan and Connell, 2001]. For *vertical selection* algorithm, we experiment with three existing methods, CORI, ReDDE and CRCS(e) ([Callan et al., 1995, Si and Callan, 2003a, Shokouhi, 2007]). For simplicity, we select the top three ranked verticals for each query.

For *item selection*, our focus is on selecting an optimal subset of documents from the documents returned from the three selected verticals. For consistency with current search engines, we return 15 documents. We implement two different ranking functions to return the documents from the verticals. We implemented two retrieval systems, a "good" (i.e. BM25) and a "bad" (i.e. a simple cosine similarity with a tf term-weighting function) ranking function. For *result presentation*, for simplicity, we are not concerned with the impact of these 15 documents on the actual presentation to users (blended presentation). Therefore, all the results are presented into a set of pre-defined fixed blocks.

In total, a combination of twelve ($6 \times 2 \times 1$) aggregated search systems are generated here. All the different combinations represent common approaches in the literature to their corresponding problem (i.e. those of vertical selection, item selection and result presentation). As stated previously, we take the top three ranked verticals to be selected for a query. To test whether the right subset of documents from each vertical have been identified, we select the top 7 documents from the first ranked vertical, top 5 documents from the second, and top 3 from the third ranked vertical.

4.4.3 Results

We used α-NDCG [Clarke et al., 2008] as a performance metric, and modelled verticals of high vertical intent as sub-topics. More details of the metric can be referred to Section 2.1.1.

We ran the 12 systems on the the collections with various levels of misclassification. We used a subset of topics that had at least two vertical intents. The average Spearman rank correlation between the performance of the systems on the three different collections is shown in Table 4.7. Note that the number of iterations of the experiments performed with respect to the larger misclassified (i.e. $> 30\%$) rate is quite small (only once)[8]. We can see that in general there is a high correlation between the systems even if misclassification increases to 30%. A random classification of documents (not shown in 4.7) leads to a moderate correlation of 0.573.

4.5 Conclusions

Our objective of this chapter is to investigate whether it is possible to create a reliable aggregated search test collection by "reusing". We describe our reusing method and have demonstrated that by identifying topics from existing test collections, a sufficient number (320) of topics with multiple vertical intents can be collected. In addition, through

[8] Although we would like to re-conduct this experiments with more iterations to gain more solid conclusions, we leave the results as they are due to the data loss of the computing server with respect to this collection.

Table 4.7: System ranking correlation for different misclassification rates

misclassified	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
correlation	0.96	0.95	0.94	0.91	0.91	0.93	0.89	0.84	0.80	0.86
iterations	5	5	5	5	5	1	1	1	1	1

simulation we have showed that aggregated search approaches can be properly evaluated even there are inherent misclassification within the verticals.

The created test collection with collected document relevance assessments in this chapter, in addition to the vertical relevance assessments collected in Chapter 3, forms an integration of data to evaluate aggregated search systems. Before we dig into how to evaluate aggregated search, we first want to understand in-depth on whether the relevance of the vertical collected by simply using vertical orientation (i.e. pre-retrieval user vertical intent) can be somewhat correlated to the vertical relevance using the document-based relevance assessments. Therefore, we further follow-up on the work presented in both Chapter 3 and this chapter by correlating both type of assessments collected.

5
Aligning Vertical Collection Relevance with User Intent

In the previous two chapters we focused on both collecting vertical relevance assessments and creating a test collection by reusing document-based relevance assessments. In this chapter, we start our investigations by studying whether both types of assessments somewhat correlate. In particular, the *user vertical intent* (collected pre-retrieval vertical orientation), the verticals the user expects to be relevant for a particular information need, might not correspond to the *vertical collection relevance*, the verticals containing the most relevant content according to the document relevance judgments collected in the test collection. It has been shown that the relevance of a vertical could depend on the relevance of the documents within the vertical collection [Gravano et al., 1994, Powell and French, 2003] and on the user's intent (orientation) to the vertical [Zhou et al., 2013a]. For evaluation purposes, from the *collection* perspective, Gravano et al. [Gravano et al., 1994] assumed that any collection (vertical) with at least one relevant document for a query is relevant. [Powell and French, 2003] refined and formalized this notion by assuming that the relevance level of the collection (vertical) depends on the number or relevant documents within. In this thesis, we call this the *collection-based vertical relevance*. On the other hand, from the *user intent* perspective, researchers [Sushmita et al., 2010, Wang et al., 2013, Zhou et al., 2013a] found that user orientation (intent), or how oriented each vertical is to a user's information need (i.e., expectation), also plays an important role in user preference of the aggregated search page. We refer to this as the *user vertical intent*. Most of the previous work either assumes that the relevance of the vertical solely depends on the collection (i.e., its recall of relevant documents) or the user intent (the user's orientation to issue the query to the given vertical search engine).

Although previous work [Zhou et al., 2013a] has shown that *user vertical intent* and *result relevance* are both correlated for influencing user experience (as we show in Chapter 3.3.1), it fails to connect both criteria within the context of evaluation in this area. The key question is whether we could align the collection-based vertical relevance with the user vertical intent for evaluation purposes.

Therefore, in this chapter [Zhou et al., 2014], we propose different approaches to define the set of relevant verticals based on document judgments The approaches differ in how they quantify the relevance of a vertical and how the ultimate set of relevant verticals is derived. We aim to address our research questions **RQ 7** to **RQ 8**, as specified in Section 1.2.1:

> **RQ 7**: Can the vertical relevance be derived from document relevance judgments and therefore ranked similarly to the user vertical intent (orientation)?
>
> **RQ 8**: Can we appropriately threshold the derived vertical rankings and ultimately align them with the binary vertical selection decision made by the users?

By conducting user studies to collect the user vertical intent (following our concluded most efficient pre-retrieval vertical orientation approach in Chapter 3), we compare those approaches on deriving collection-based vertical relevance and investigate which approach best aligns with the actual user intent. Following the data collected, we correlate the collection-based relevant verticals obtained from these approaches to the real user vertical intent, and show that they can be aligned relatively well. We study this in the context of over a hundred heterogeneous resources (search engines). The scale and diversity of the resources used has not been studied previously for our task. The set of relevant verticals defined by those approaches could therefore serve as an approximate but reliable ground-truth for evaluating vertical selection, avoiding the need for collecting explicit user vertical intent.

The main elements of this chapter are summarized in Table 5.1. We first describe different approaches to obtain both the vertical ranking and a set of relevant verticals using the collection-based document judgments (Section 5.1). We then conduct a user study (Section 5.2) to obtain the vertical ranking and relevant vertical sets from the user (as the ground-truths). Finally, we evaluate different approaches proposed in Section 5.1 and study how well they can be aligned with the user intent (Section 5.3), after which the chapter is concluded (Section 5.4).

5.1 Vertical Collection Relevance

Formally, given a set of verticals $V = \{v_1, v_2, ...v_n\}$, the collection-based vertical relevance I_t^C derived from the collection C for topic t is represented by a weighted vector $I_t^C = \{i_1, i_2, ...i_n\}$, where each value i_k indicates the relevance of the given vertical v_k to topic t. A vertical could contain multiple resources (search engines). For example, an "image" vertical could contain resources such as Flickr and Picasa. Therefore, each vertical v_i consists of a set of resources $v_i = \{r_1, r_2, ...r_m\}$ while each resource r_j consists of a set of documents $r_j = \{d_1, d_2, ...d_k\}$. Given all the relevance judgments $rel(d_l, t)$ between any document d_l and a topic t, we aim to derive I_t^C.

Ultimately, given the collection-based vertical relevance I_t^C, we aim to threshold it in order to obtain the final binary verticle relevance vector $S_t^C = \{s_1, s_2, ...s_n\}$ where each value s_k is either 1 (indicating corresponding vertical v_k is relevant and should be selected in VS) or 0 (indicating irrelevant and should not be selected).

Table 5.1: Summary of a Variety of Approaches to Analyze Collection-based Vertical Relevance Against User Vertical Intent.

Tasks	Ranking of Verticals		Set of Relevant Verticals	
Variants	(1). Resource Relevance	(2). Aggregation	(3). Vertical Dependency	(4). Thresholding Criteria
Collection	a. **K** (Key Rel Doc Recall) b. **G** (Doc Graded Precision)	a. **MR** (Maximal Resource) b. **AR** (Average of Resource)	a. **D** (Dependent) b. **I** (Independent)	a. **I** (Individual vertical utility) b. **O** (Overall vertical set utility)
User Study	Fraction of majority user preference of each vertical over “General Web”		User type: 1. risk-seeking (diversity); 2. risk-medium; 3. risk-averse (relevance)	
Evaluation	Spearman Correlation		Precision, Recall and F-measure	

5.1.1 Approaches

We describe approaches to derive the vertical ranking, followed by methods to infer the set of relevant verticals.

Vertical Ranking

The strategies to derive the collection-based vertical relevance I_t^C (i.e., the vertical score) from the document relevance judgments $rel(d_l, t)$ vary in two aspects: (1). **(Resource Relevance)** the way to estimate the relevance of each resource within a given vertical; and (2). **(Vertical Relevance Aggregation)** the way to aggregate scores of the resources within the vertical to derive the vertical score.

Following previous work [Demeester et al., 2013, Gravano et al., 1994], we propose two approaches to estimate **resource relevance**: (a). **K** (Key): using the recall of "key" (most relevant) documents in the resource and (b). **G** (Graded precision): the graded precision of documents in the resource. The **K** approach is similar to the assumption made in Gravano et al [Gravano et al., 1994] and using "key" is to reflect the relevance of the resource to return the most relevant results that maintain high impacts on user search experience [Sanderson et al., 2010]. The **G** approach is following the evaluation setup [Demeester et al., 2013] made in the TREC FedWeb track 2013[1] and the essential idea is to characterize the effectiveness of each resource to "recall" relevant documents in a similar fashion as in previous work [Powell and French, 2003] when graded relevance judgments are available. The graded precision is proposed in [Kekäläinen and Järvelin, 2002] and the relevance of a resource for a given query is determined by calculating the graded precision on the top 10 results. This takes the graded relevance levels of the documents in the top 10 into account, but not the ranking. The relevance levels are taken from the TREC Web track. The following weights are given to the relevance levels of documents (Nav: 1, Key: 1, Hrel: 0.5, Rel: 0.25, Non relevant: 0). For example, a resource that returns 1 Key and 2 Relevant pages in its top 10 has a graded precision of $(1 + 2 * 0.25)/10 = 0.15$.

Then given the estimated resource relevance scores, we test two ways to **aggregate** those scores in order to obtain the vertical scores (rankings) I_t^C: (a). **(MR)** Maximal Resource score and (b). **(AR)** Average Resource score. The **MR** approach reflects most of current web search setting that one vertical solely contains one best performing resource while the **AR** approach represents the averaged vertical performance.

By combining the different *resource relevance* and *aggregation* methods, we obtain four approaches to quantify I_t^C: **KMR**, **KAR**, **GMR**, **GAR**. Since we could also apply the same technique to the whole vertical (rather than resource), we propose another approach **GV** by using graded precision on all the documents within the entire vertical[2].

Relevant Vertical Set

To infer the set of relevant verticals S_t^C from the obtained collection-based vertical relevance I_t^C, we argue that the strategies could vary in two aspects: (3). (**Vertical De-**

[1] https://sites.google.com/site/trecfedweb/2013-track.

[2] We do not propose **KV** approach since practically, it outputs the same vertical ranking to **KAR** approach.

pendency): assumption of whether vertical relevance is dependent on each other; and (4). (**Thresholding Criterion**): assumption of which is the criterion of thresholding. For (3). **(Vertical Dependency)**, we tested both assumptions. By assuming (a). **(D)** Dependent, we normalize the vertical scores across all verticals following previous work [Arguello et al., 2009]. By assuming (b). **(I)** Independent, we simply use the original vertical scores I_t^C.

For (4). **(Thresholding Criterion)**, we tested two different approaches. The differences between them are the criterion that the thresholding is based on: (a). **(I)** Individual vertical score: the individual vertical relevance scores; (b). **(O)** Overall relevance of the vertical set. The **I** approach basically assumes that the individual vertical requires a certain relevance to remain in the relevant vertical set while the **O** approach assumes that the relevant vertical set is required to maintain a certain percentage of relevance of the whole vertical set.

By combining *vertical dependency* and *thresholding criterion*, we obtain four different approaches to infer S_t^C from I_t^C: **DI**, **DO**, **II**, **IO**.

5.1.2 TREC FedWeb'13 Data

In this study, we use the TREC 2013 FedWeb track data [Demeester et al., 2013]. The dataset contains 50 test topics and 157 crawled resources. It also categorizes the resources into different verticals and provides a set of 24 verticals, as shown in Table 5.2. Each vertical consists of a set of resources (search engines). For each resource, the top 10 retrieved document results are returned. The relevance judgments are made on each document with five graded relevance levels: Non (not relevant), Rel (minimal relevance), HRel (highly relevant), Key (top relevance), and Nav (navigational).

The 50 test topics were chosen in such a way to avoid a strong bias towards general web search engines. For the most important verticals (in terms of number or size of resources, e.g. Video, Blogs), many topics provide a significant number of relevant results. In addition, at least a few topics targeting smaller verticals (e.g., Recipes, Travel) are also selected.

5.2 Vertical User Intent Study

Given a set of verticals $V = \{v_1, v_2, ...v_n\}$, the vertical user intent I_t^U for topic t is represented by a weighted vector $I_t^U = \{i_1, i_2, ...i_n\}$, where each value i_k indicates the relevance of the given vertical v_k to topic t. To obtain I_t^U, we conducted a user study, asking assessors $U = \{u_1, u_2, ..., u_m\}$ to make binary decisions over all verticals V: $A = \{a_1, a_2, ..., a_n\}$. Therefore, we have a $m \times n$ matrix M_t for topic t. We aim to derive I_t^U by aggregating M_t and ultimately obtain a binary vector indicating the set of relevant verticals $S_t^U = \{s_1, s_2, ...s_n\}$ where each value s_k is either 1 or 0.

We conducted this user study following our work in Chapter 3 on gathering user vertical intent [Zhou et al., 2013a, 2012a]. Basically, two assumptions were made in guiding the assessment. Firstly, instead of asking assessors to associate an absolute score to each vertical, we asked them to make *pairwise preference assessments*, comparing each vertical in turn to the reference "general web" vertical (i.e. "is adding results from

Table 5.2: 24 Verticals Used in FedWeb'13: a Vertical Consists of a Set of Resources (Search Engines), Each Retrieving One Unique Type of Documents.

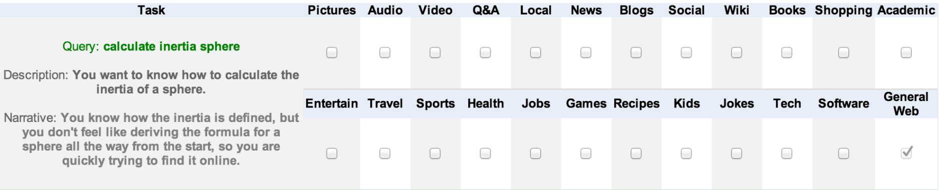
Vertical	Document	Resource Count	Type
Pictures	online pictures	13	media
Audio	online audios	6	
Video	online videos	14	
Q&A	answers to questions	7	genre
Local	local information pages	1	
News	news articles	15	
Blogs	blog articles	4	
Social	social network pages	3	
Encyclopedia	encyclopedic entries	5	
Books	book review page	5	
Shopping	product shopping page	9	
Academic	research technical report	18	
Entertainment	entertainment pages	4	topic
Travel	travel pages	2	
Sports	sports pages	9	
Health	health related pages	12	
Jobs	job posts	5	
Games	electronic game pages	6	
Recipes	recipe page	5	
Kids	cartoon pages	10	
Jokes	joke threads	2	
Tech	technology pages	8	
Software	software downloading pages	3	
General Web	standard web pages	6	

this vertical likely to improve the quality of the *ten blue links*?”). Secondly, instead of providing actual vertical results to the assessors, we only provided the *vertical names* (with a description of their characteristics presented before their assessments). Although this may not be ideal from an end-user perspective (as different assessors might have different views on the perceived usefulness of a vertical, especially as the vertical items are hidden), this assumption helped to lower the assessment burden, and yet reflects the perceived vertical user intent (orientation).

We used the same FedWeb’13 data as described in Section 5.1.2. Most of the assessors are university students who were recruited to participate via a web interface. Most of the participants are university students who use English everyday and they are paid to make the assessments. To eliminate order bias, we randomized all topics into a set of pages (with five topics per page) and provided each assessor the option to assess as many pages as he/she wished. A screenshot of one examplar task is presented in Figure 5.1.The screen shot in the Figure shows one task (out of five on the screen) that the participants saw. The participants can navigate to the next five topics by pushing the ”I want to assess five more” button at the bottom of the screen. There are additional instructions given to the assessors. The first user study page introduces this user study and collects participant information. The second page describes the unique characteristics of each vertical and search results from each vertical results for one exemplar information need. The third page shows one example of how to make the judgments. Then the latter pages aim to collect assessments as shown in Figure 5.1.

In total, we collected 20 assessment sessions (i.e., assessors) with a total of 845 assessments. The average number of relevant verticals per topic and per session is 2.64, with a standard deviation of 1.28. Similar to previous findings [Zhou et al., 2012a], the mean of inter-annotator Fleiss’ Kappa [Fleiss, 1971] is *moderate* (0.48), showing that assessors might have different preferences over the relevance of verticals, despite the clearly described query information need (as seen from the description and narrative shown in Figure 5.1).

To derive I_t^U, we use the fraction of majority user preferences for each vertical v_k over “General Web” as the vertical score i_k. To further obtain S_t^U, we threshold the majority user preference for each vertical i_k in I_t^U. It has been shown in our data (moderate inter-annotator Fleiss’ Kappa agreement) and previous work [Zhou et al., 2012b] that the user’s preferred number of verticals varies significantly and different users tend to have different risk-levels. By thresholding 30%, 60% and 90% of majority user preference, we obtain three different types of S_t^U, representing three types of users respectively: *risk-*



Task	Pictures	Audio	Video	Q&A	Local	News	Blogs	Social	Wiki	Books	Shopping	Academic
Query: calculate inertia sphere Description: You want to know how to calculate the inertia of a sphere. Narrative: You know how the inertia is defined, but you don't feel like deriving the formula for a sphere all the way from the start, so you are quickly trying to find it online.	□	□	□	□	□	□	□	□	□	□	□	□
	Entertain	Travel	Sports	Health	Jobs	Games	Recipes	Kids	Jokes	Tech	Software	General Web
	□	□	□	□	□	□	□	□	□	□	□	☑

Figure 5.1: The Screenshot of An Examplar Task of the User Intent Study Following Our Conclusion in Chapter 3.

seeking, *risk-medium* and *risk-averse*. The *risk-seeking* users prefer *diversity* of verticals presented (with a mean of 3.08 relevant verticals) while the *risk-averse* users are more careful when selecting verticals (with a mean of 0.52 relevant verticals): they only select verticals (as relevant) when highly confident (large fraction of user's preferences). The *risk-medium* is an average user, with a mean of 1.68 relevant verticals (following a similar distribution as shown in [Arguello et al., 2009]).

5.3 Evaluation

We evaluate the alignment between collection and user, on both vertical rankings and ultimate relevant vertical sets.

5.3.1 Vertical Ranking

Given the collection-based vertical relevance I_t^C derived from document relevance judgments and user vertical intent I_t^U obtained from user preference judgments, we evaluate whether they align with each other. We aim to evaluate five different approaches of utilizing relevance judgments for ranking verticals, as described in Sec 2.1.1. Specifically, we utilize the nonparametric Spearman Rank Correlation Coefficient as our main metric to measure the correlation between a collection-based vertical ranking and a user-based one. Since we are more concerned with highly ranked verticals (potentially relevant), we also investigate whether there are overlaps between the top-3 and top-5 ranked verticals in the collection-based and user-based rankings. The evaluation results of different approaches are shown in Table 5.3.

Several trends can be observed. Firstly, all the collection-based approaches have a moderate correlation (0.6-0.7) with the user-based vertical ranking. We also study whether this correlation is statistically significant (against random) by performing a permutation test. We found that the correlation for all the five approaches are statistically significant (with p <0.05). Note that the performance difference between different approaches is marginal while the approaches using the graded precision metric outperforms the others. Secondly, we observed that there tends to be some overlap between the top ranked verticals from both collection-based and user-based vertical rankings, albeit moderate (0.4-0.5). However, it is interesting to see that when using simple metrics on document-based relevance judgements, around half of the top-ranked verticals are

Table 5.3: Spearman correlation and overlap of top-k verticals between vertical rankings from collection-based vertical relevance I^C and vertical user intent I^U.

Approaches	KMR	KAR	GMR	GAR	GV
Correlation	0.656	0.659	0.664	0.689	0.671
Overlap (5)	0.496	0.492	0.516	0.532	0.508
Overlap (3)	0.340	0.340	0.327	0.400	0.353

Table 5.4: Precision, Recall and F-measure of the set of relevant verticals using collection-based vertical relevance I^C from GMR approach against user vertical intent I^U with different risk-level.

User Intent	Risk-seeking (diversity)			
Approaches	DI	DO	II	IO
Precision	0.264	0.289	**0.338**	0.271
Recall	**0.829**	0.803	0.796	0.782
F-measure	0.380	0.406	**0.446**	0.384

User Intent	Risk-medium			
Approaches	DI	DO	II	IO
Precision	0.303	0.330	**0.391**	0.327
Recall	**0.552**	0.523	0.516	0.520
F-measure	0.367	0.387	**0.413**	0.383

User Intent	Risk-averse (relevance)			
Approaches	DI	DO	II	IO
Precision	0.339	0.377	**0.421**	0.381
Recall	**0.426**	0.398	0.401	0.394
F-measure	0.357	0.374	**0.379**	0.373

aligned with the user intent.

In summary, our experiments suggest that collection-based vertical relevance can be utilized as an approximate surrogate for measuring user's vertical intent, and vice versa.

5.3.2 Vertical Relevant Set

We study whether the obtained set of relevant verticals after thresholding is aligned with the ones derived from the user perspectives. For simplicity, we only present results on thresholding with one collection-based vertical ranking approach **GMR** (Graded precision of Maximal Resource) since we found similar results across all those different approaches.

As we have mentioned, we defined three types of ground-truths, representing three different types of users: risk-seeking users prefer a large set of diverse verticals, while the risk-averse users prefer selecting verticals only when they are most relevant, and risk-medium are in between. We test different thresholding approaches for these three user settings.

For each thresholding approach, its numerical threshold was determined based on iterative data analysis, such that the maximum number of relevant verticals for any test topic could not exceed five. In addition, since almost all the FedWeb'13 test topics target verticals, we also make sure that at least 80% of the topics have at least one relevant vertical using the selected threshold.

The results are shown in Table 5.4. We observe similar performance trends for different thresholding approaches under different user settings (risk-level). **II** (Independent Individual) thresholding approach performs best in terms of precision and F-measure while **DI** (Dependent Individual) thresholding approach generally would achieve better recall. Generally, an F-measure of around 0.4 could be achieved by mapping the estimated collection-based relevant vertical set with the users' relevant (intended) vertical set. Although not particularly high, this still shows that vertical collection relevance could be aligned relatively well with users' vertical intent and therefore this could serve as a surrogate of ground-truths for evaluating vertical selection.

5.4 Conclusions

This chapter focused primarily on analyzing the relationship between the vertical-level and document-level relevance assessments. In this chapter, we propose a set of different approaches to utilize document judgments to derive the set of relevant verticals. We evaluate the effectiveness of those approaches by correlating with the user vertical intent obtained from a user study. We found that collection-based vertical relevance can be aligned relatively well with users' vertical intent. This implies that we could reliably use document relevance judgments to evaluate vertical selection for capturing user intent in heterogeneous federated web search.

The alignment between collection-based relevant verticals and user vertical intents has moderate (and significant) correlation. For the conclusion, our recommendation is not to infer vertical intent by using collection-based judgments. Rather, due to the reasonable alignment, we conclude that we might be able to use collection-based judgments as the approximate ground-truth to evaluate vertical selection (and vice versa).

Here we summarize the assessment part of this thesis (Part II). By conducting several user studies and creating a test collection by reusing, we collect both the vertical-level and document-level relevance assessments. In addition, we analyze the relationship between them and found that collection-based relevant verticals and user vertical intents has moderate but significant correlation. We conclude from the above studies, we now have a better understanding of the relevance from the perspective of vertical. We hypothesize that these assessments available enables more accurate modelling of the user behavior on aggregated search pages and we would be able to use those to evaluate key aggregated search components such as vertical selection, item selection and result presentation, or the entire aggregated search system. Therefore, with the test collection available, following up on this hypothesis, we aim to move on to the evaluation metric part of Cranfield Paradigm for aggregated search (Part III).

Part III

On the Evaluation Metrics of Aggregated Search

6 Evaluating Reward and Risk for Vertical Selection

Part II of this thesis we described above aims to understand better on the assessments for Cranfield paradigm for aggregated search. We now turn to another component of Cranfield paradigm, i.e. the evaluation metrics. Basically, we aim to propose a set of aggregated search metrics that utilize the assessments available, in order to model the user behavior and enable reliable and trustworthy evaluation of aggregated search.

In this chapter, we start our investigations by studying evaluating one key component of aggregated search: vertical selection. Generally, it is not always the case that provide additional results from other verticals can benefit the users. Only selecting relevant verticals that a large population of users that are favoring can be rewarding while selecting irrelevant verticals that not a lot of users intended can result in the risks of hurting the user experience. This chapter addresses our research questions **RQ 9** to **RQ 10**, as specified in Section 1.2.1. We propose the risk-aware vertical selection metrics that aims to study a number of vertical selection approaches with respect to this. We aim to answer:

> **RQ 9**: For evaluating vertical selection, rather than solely consider reward (selecting relevant verticals), can we measure the performance on maximising reward while minimising risk (selecting irrelevant verticals)?
>
> **RQ 10**: How effective and robust are existing vertical selection approaches considering the varying types of user (risk-averse and risk-seeking)?

As we mentioned, when selecting suitable verticals, there exists the potential to both help (selecting relevant verticals) and harm (selecting irrelevant verticals) the existing result set. A VS system should only select a vertical when it is confident that it will benefit most users while seldom frustrating others. Existing work evaluates VS based solely on maximising reward (the number of queries correctly classified as relating to a vertical [Arguello et al., 2009]), or the average correlation with the "perfectly ranked" reference page [Arguello et al., 2011b]. We argue that for VS, reward must be considered in conjunction with risk. We argue that maximising the reward alone is not sufficient, and that a robust VS approach and its evaluation should focus on maximising reward while minimising risk.

We propose a new risk-aware VS evaluation metric. Rather than treating a vertical as either relevant or irrelevant given a query, as mostly done in current work [Arguello et al., 2009], we propose a general framework to evaluate the reward and risk for VS on a per user basis. This is motivated by the fact that current research [Zhou et al., 2012a] shows that the level of inter-annotator agreement for what constitutes a ‘relevant’ vertical is low (users’ preferred verticals are diverse). Our proposed metric is flexible as it allows systems to be evaluated across a population of users, where users may have varying levels of risk (risk-averse vs. risk-seeking) and may have varying preferences across verticals (vertical relevance is user specific). In this chapter, we perform an analysis of the effectiveness of different VS approaches across these different types of user [Carterette et al., 2011]. Furthermore, we present an analysis of the robustness of VS approaches across all users with various levels of risk[1].

We treat VS as a multi-label classification problem (multiple verticals are relevant to a query) and we train a set of VS systems according to different controlled risk-level (some systems are more risk-averse than others). We then analyse these trained VS systems with varying types of user (risk-averse and risk-seeking). We hypothesise that:

- (**effectiveness**) some VS approaches are better suited to some types of users than others;
- (**robustness**) some VS approaches are more robust for a mixture of varying types of users than others.

The remainder of this chapter is organized as follows. Section 6.1 outlines our proposed risk-aware VS metric. In Section 6.2, we formally describe the problem of multi-label vertical classification and list the features used. In Section 6.3, we empirically evaluate the effectiveness and robustness of those approaches using our proposed risk-aware metric. We conclude this chapter in Section 6.4.

6.1 Evaluating Reward and Risk

We present our risk-aware metric for VS, which considers an entire population of users’ vertical preferences for a query.

6.1.1 Problem Formulation

Let $V = \{v_1, v_2, ...v_n\}$ be a set of verticals that can be selected to present along with “general web” results W, for a given query $q \in Q$. Let $V_q^{u_i}$ be a set of verticals that a user $u_i \in U$ would like to see in the result set with “general web” results for query q. These user-specific assessments can be obtained by either conducting a user study that explicitly asks users for their preferences [Zhou et al., 2012a] or be estimated by mining query logs [Ponnuswami et al., 2011b]. We model this subjective view of vertical relevance where users’ vertical preferences can be different [Zhou et al., 2012a]. Therefore, $V_q^{u_i} \subset V$ and $V_q^{u_j} \subset V$.

Furthermore, assume a vertical selection system s_j selects a vertical set $V_q^{s_j}$ for q. Then, for a specific user u_i, the utility of vertical search system s_j is based on both

[1] An analysis of the distribution of risk-levels in the user population lies outside the scope of this work. This information could be estimated from query logs or through a survey of a sample of users.

reward and *risk*. *Reward* is related to the number of verticals selected by s_j that user u_i deems relevant ($V_q^{s_j} \bigcap V_q^{u_i}$). While *risk* is related to the number of verticals selected by s_j that user u_i deems non-relevant ($V_q^{s_j} \bigcap \ (V - V_q^{u_i})$).

Furthermore, each user has his/her own estimated trade-off between reward and risk. For example, one user might be *risk-seeking* and prefers to have a page with some relevant verticals but does not mind viewing many non-relevant ones. On the contrary, another users might be *risk-averse* and prefers the page to only contain relevant verticals. Therefore, the main aim of the proposed metric is to model the trade-off between so-called *reward* and *risk* for each user u_i.

6.1.2 Risk-aware Metric

For a given user u_i and system s_j that returns $V_q^{s_j}$, we define the *reward* and *risk* as user-specific vertical $recall$ and vertical $fallout$ respectively as follows:

$$reward_q^{u_i}(V_q^{s_j}) = \frac{|V_q^{s_j} \bigcap V_q^{u_i}|}{|V_q^{u_i}|} \tag{6.1}$$

$$risk_q^{u_i}(V_q^{s_j}) = \frac{|V_q^{s_j} \bigcap \ (V - V_q^{u_i})|}{|\bar{V}_q^{u_i}|} \tag{6.2}$$

To combine the above measure and also incorporate the user's trade-off between reward and risk, we model the metric as a linear combination of reward and risk:

$$util(V_q^{s_j}, \alpha_q^{u_i}) = (1 - \alpha_q^{u_i}) \cdot reward_q^{u_i}(V_q^{s_j}) + \alpha_q^{u_i} \cdot (1 - risk_q^{u_i}(V_q^{s_j})) \tag{6.3}$$

where $\alpha_q^{u_i}$ is a user-specific parameter that controls the trade-off between reward and risk. Setting $\alpha_q^{u_i} = 1$ leads to a risk-averse metric where returning zero irrelevant verticals would be optimal, while setting $\alpha_q^{u_i} = 0$ leads to a risk-seeking metric where returning as many relevant verticals would be optimal.

The utility of the system s_j is averaged over all $q \in Q$, and within each q is averaged over all users U. We define the utility of the system as follows:

$$Util(s_j, \alpha) = \frac{1}{|Q|} \sum_{q \in Q} \frac{\sum_{u_i \in U} util(V_q^{s_j}, \alpha_q^{u_i})}{|U|} \tag{6.4}$$

This $Util(s_j, \alpha)$ function treats all (both popular and long-tailed) queries equally and is not biased to popular queries. Although other approaches to derive utility within this framework are possible, we will leave them for future work.

At this point we have one utility metric for evaluating a VS system, accounting for reward and risk. The metric depends on the user-specific and query-specific reward-risk tradeoff parameter $\alpha_q^{u_i}$, which we need to set. In this chapter, we assume that for each query q, users have the same trade-off level (α) between reward and risk. Furthermore, we assume a uniform distribution of $\alpha_q^{u_i}$ across all users. We leave the work of discovering the distribution of *risk-seeking* and *risk-averse* for future work. Using our metric we can compare vertical selection approaches for both *risk-seeking* and *risk-averse* users over a set of queries Q. Furthermore, we can measure the robustness of the VS approach over all types of users (assuming uniformity) by iterating over all values of α for all queries in Q.

6.2 Multi-label Classification

We introduce the risk-aware multi-label classification approach, followed by detailed descriptions of features used.

6.2.1 Risk-aware Classification Approach

The approach to classification consists of two phases: testing and training. We separate 56 queries (conforming to a real-world distribution of verticals [Arguello et al., 2009]) as a training set. This is used for determining a threshold γ (see below). We use the remaining dataset (264 queries) for testing the approaches.

We use a thresholding approach to select verticals. For a set of verticals $V = \{v_1, v_2, ...v_n\}$ with scores $X^{s_j} = \{x_1, x_2, ...x_n\}$ (generated by a vertical selection approach s_j) and a threshold γ, we denote $V^{s_j}_{x_i > \gamma}$ as the set of verticals with each vertical v_i whose score $x_i > \gamma$. If no vertical has $x_i > \gamma$, then $V^{s_j}_{x_i > \gamma} = \emptyset$. Note that each vertical score x_i is obtained by normalising across all vertical scores.

In essence, the vertical scoring functions of each VS approach is adapted to multi-label vertical selection by selecting the top-k verticals where k is decided by a threshold γ. The threshold is trained on the training set. If no verticals receive a score greater than the threshold, no verticals are deemed relevant for that query.

With respect to the risk-aware training, for a given vertical selection approach s_j with scores over all verticals $X^{s_j} = \{x_1, x_2, ...x_n\}$, we train a set of systems $S_j = \{s^{\alpha_1}_j, s^{\alpha_2}_j, ...s^{\alpha_m}_j\}$ where each system varies in its reward-risk trade-off operating point (by setting different training objective functions with different α, and obtaining corresponding γ), i.e. some of the systems are trained to be more risk-averse whereas others to be more risk-seeking. The optimal threshold γ^* for a given system s^{α}_j (with reward-risk trade-off α) is trained as follows:

$$\gamma^* = argmax_{\gamma}\ Util(V^{s^{\alpha}_j}_{x_i > \gamma}, \alpha) \tag{6.5}$$

Therefore, for each feature (vertical selection approach s_j), we iterate α and obtain a set of systems S_j.

6.2.2 Features

We investigate a number of resource selection approaches (CORI [Callan et al., 1995], Clarity [Cronen-Townsend et al., 2002], GAVG [Seo and Croft, 2008], ReDDE [Si and Callan, 2003a], CRCS(l) [Shokouhi, 2007], CRCS(e) [Shokouhi, 2007]) as features for multi-label VS approaches. We use each feature individually for training and aim to compare them. While these approaches derive evidence from the same source (sampled vertical representation), they model different aspects of the sources under consideration. CORI, Clarity and GAVG model the similarity between the query and the source, whereas ReDDE, CRCS(l) and CRCS(e) model the collection's average document score in a full-dataset retrieval (all sources together). As follows, we briefly discuss each resource selection approaches utilized and more details can be referred in Section 2.2.1.

CORI

CORI adapts INQUERY's inference net document ranking approach to collection. Here, all statistics are derived from sampled documents rather than the full collection.

Clarity

Clarity is a retrieval effectiveness prediction algorithm that measures the similarity between the language of the top ranked documents and the language of the collection, estimated using the Kullback-Leibler divergence between the query θ_q and the collection language model θ_{v_i}.

$$Clarity_q(v_i) = \sum_{w \in v_i} P(w|\theta_q) log_2 \frac{P(w|\theta_q)}{P(w|\theta_{v_i})} \tag{6.6}$$

Geometric Average

GAVG issues the query to a centralized sample index, one that combines document samples from every vertical, and scores vertical v_i by the geometric average query likelihood from its top m sampled documents.

$$GAVG_q(v_i) = \left(\prod_{d \in topm} P(q|\theta_d) \right)^{\frac{1}{m}} \tag{6.7}$$

ReDDE

ReDDE scores a target collection based on its expected number documents relevant to the query. It derives this expectation from a retrieval of an index that combines documents sampled from every target collection. Given this retrieval, *ReDDE* accumulates a collection score $ReDDE_q(v_i)$ from its document scores $P(q|\theta_d)$, taking into account the difference between the size of the original collection N^{v_i} and a sampled set size N^{samp}.

$$ReDDE_q(v_i) = \frac{N^{v_i}}{N^{samp}} \sum_{d \in topm} I(d \in v_i) P(q|\theta_d) \tag{6.8}$$

where $I(.)$ is a indicator function.

CRCS

Like *ReDDE*, *CRCS* issues the query to a centralized sample index and scores a collection according to an accumulation of a more refined estimation of document score. Specifically, the document score for *CRCS(l)* and *CRCS(e)* are estimated by a linear or a negative exponential weighting according to its presented position respectively.

$$CRCS(l)_q(v_i) = \frac{N^{v_i}}{N^{samp}} \sum_{d_j \in topm} (m - j) \tag{6.9}$$

$$CRCS(e)_q(v_i) = \frac{N^{v_i}}{N^{samp}} \sum_{d_j \in topm} \alpha \cdot exp(-\beta \cdot j) \tag{6.10}$$

where $\alpha = 1.2$ and $\beta = 2.8$ in our setting.

Table 6.1: Distribution of Number of Queries Assigned to Majority User Prefered Verticals (In Total 320 Queries)

Verticals	Image	Video	Recipe	News	Book	Blog
Query num	41	13	7	22	25	22
Verticals	Answer	Shopping	Discuss	Scholar	Wiki	Web-only
Query num	38	4	38	11	139	141

6.3 Experiments

Our experiments aim to investigate various resource selection approaches under our risk-aware multi-label classification framework. We report the data used in the experiments first, followed by the main experimental results on both *effectiveness* and *robustness*.

6.3.1 Data

The user-specific preferred vertical ground-truth information of each query ($V_q^{u_i}$) is obtained by only providing the vertical names (with a description of their characteristics) and asking a set of assessors to make pairwise preference assessments, comparing each vertical in turn to the reference "general web" vertical [Zhou et al., 2012a]. This is the data that we collected similar to the approach described in Section 5.2 using the test collected we created by reusing in Section 4.3. We used an existing web test collection described in Section 4.3 [Zhou et al., 2011] to obtain the vertical representations used for the vertical selection approaches. The verticals used and the distribution of majority user preferred verticals (more than 50% of the users preferred the vertical to "general web") for all queries for the collection are described in Table 6.1.

6.3.2 Evaluating VS Approaches

Effectiveness

A VS approach s_j trained on a given user risk-level α is tested on the corresponding type of user (with same α). An approach is *effective* if prediction of relevant verticals $V_q^{s_j}$ can satisfy users of that type (i.e. high $Util(s_j, \alpha)$).

The main evaluation results on effectiveness for single-feature (each resource selection approach) classifier runs are shown in Figure 6.1. When only reward is considered ($\alpha = 0$), all of the approaches perform comparably. However, when risk is considered ($\alpha > 0$), we observe that in general, ReDDE performs consistently better than the other approaches. From a 2-tailed paired t-test ($p < 0.05$), we find that ReDDE is significantly better than GAVG and CRCS(e) at $\alpha = 0.3, 0.4$, CRCS(l) at $\alpha = 0.3$, Clarity and CORI at $\alpha = 0.2, 0.3, 0.4, 0.5$. Of the VS approaches tested CRCS(l) and ReDDE are more risk-aware (when $\alpha > 0.4$ for example). However, when favouring reward (low α), GAVG and ReDDE achieve higher results. CORI and Clarity are, on average, the worst approaches across many values of α.

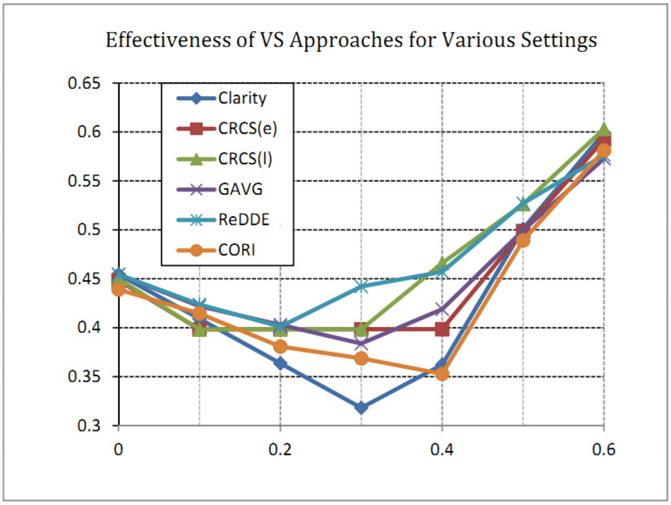


Figure 6.1: Comparing Effectiveness for Various Vertical Selection Approaches using Risk-Aware Vertical Selection Metric

We also empirically observe that different approaches perform differently for a range of queries whereas some of them hinder/increase the performance of more queries than the other when applying vertical selection. The percentage of benefited and hindered queries conforms to the training setting of the reward-risk trade-off. This demonstrates the need for current VS approaches to be more risk-aware.

In conclusion, comparably, ReDDE and CRCS(l) achieve the best performance on *effectiveness* in those settings, mostly with a large range of queries benefited and a small amount hindered.

Robustness

Rather than evaluating on one single type of user, robustness of VS approach is measured over all types of users (assuming uniformity) by iterating over risk-level α for all queries.

The main evaluation results on robustness are shown in Figure 6.2. Firstly, we can observe a general trend that VS approaches that balance the trade-off between reward and risk perform better than the ones that considers solely reward or solely risk. This is not surprising since VS approaches that solely maximise reward frustrate most of users that are *risk-averse*. On the contrary, only minimising risk could degrade user experience for users that are *risk-seeking*.

Secondly, it can be observed that in general, CRCS(l) perform more robust than other approaches. From a 2-tailed paired t-test ($p < 0.05$), we find that CRCS(l) is significantly more robust than all other approaches when $\alpha >= 0.4$. When $\alpha < 0.4$, CRCS(l) is significantly better than CORI at $\alpha = 0.0, 0.1$, GAVG, Clarity, CRCS(e) and Clarity at $\alpha = 0.2, 0.3$, ReDDE at $\alpha = 0.2$. Of the VS approaches tested, CRCS(l) and Clarity are more risk-aware (when $\alpha > 0.4$ for example). However, when favouring reward (low α), GAVG and ReDDE achieve higher results. CORI and CRCS(e) are, on average, the worst approaches across many values of α. We can conclude that CRCS(l) achieve the best performance on *robustness* in our settings.

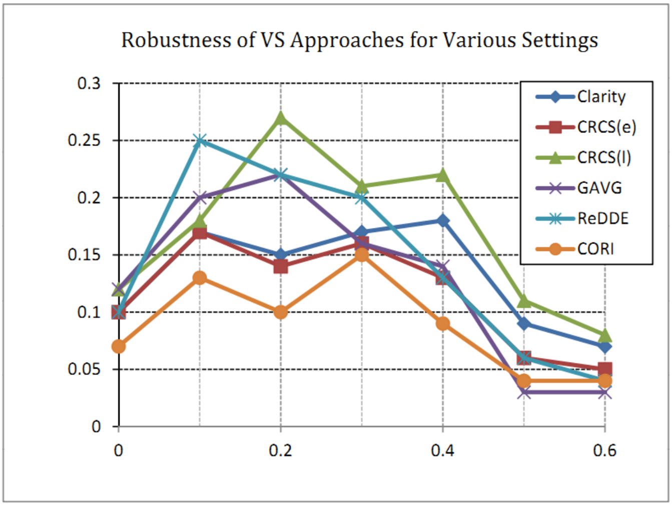


Figure 6.2: Comparing Robustness for Various Vertical Selection Approaches using Risk-Aware Vertical Selection Metric

6.4 Conclusions

This chapter incorporates a risk-aware evaluation of vertical selection approaches in a multi-label classification framework. We propose a novel multi-label vertical selection evaluation metric that incorporates both rewards and risks. We present a detailed empirical analysis of both effectiveness and robustness of current vertical selection approaches. We demonstrate that ReDDE is the most effective VS approach and CRCS(l) is the most robust.

Although evaluating vertical selection could provide lots of insights on the performance of aggregated search systems, however, in order to further understand and evaluate item selection and result presentation, the other two important components of aggregated search, we need an evaluation framework that can incorporates all the key components of aggregated search for the full evaluation. Therefore, we further follow-up on the work presented in this chapter by proposing a general utility-effort framework to evaluate the ultimate aggregated search pages in the next chapter.

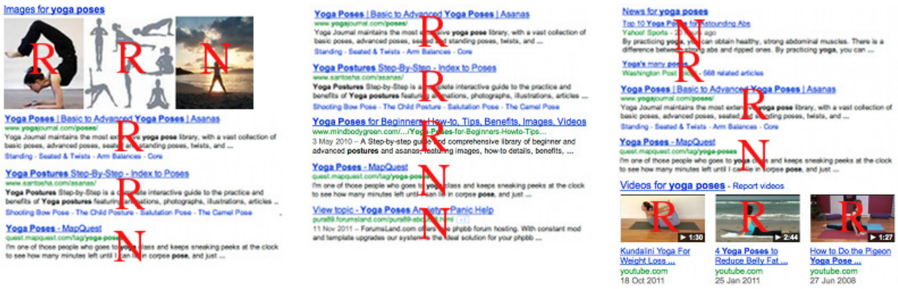
7 Evaluating Aggregated Search Pages

In the previous chapter we focused on evaluating vertical selection that considers both risk and reward. In this chapter, we focus on proposing evaluation metrics to measure aggregated search pages. Although various approaches exist for selecting relevant verticals or optimising the aggregated search result page, evaluating the quality of an aggregated page is an open question. Consider the query "yoga poses" which suggests that a visual element in the result page would be useful to many users. Furthermore consider that 75% of users who issue this query would prefer "image" results, 60% would prefer "video" results, and 10% would prefer "news" results, to "general web" results. Figure 7.1 shows three possible aggregated search pages[1] (A, B, and C) for the sample query. It is clearly difficult to objectively ascertain the aggregated search page that represents a more effective returned set, as there are a variety of compounding factors that could affect a user preference. A user may prefer a page because of his/her preference towards a specific vertical (vertical preference). In such a case, a user may prefer page A because it contains more images. A user who prefers a result set with more items that are topically relevant might prefer page C, whereas a user who prefers more relevant items towards the top of the page (presentation preference) might prefer page B. Furthermore, a user who desires a more diverse returned set (vertical diversity) may prefer page C. Any combination of those factors can influence the perceived quality and user preference of the pages.

Following the key component evaluation, we turn to a thorough evaluation the entire aggregated search system, i.e. measuring the effectiveness of the ultimate aggregated search pages. This chapter addresses our research questions **RQ 11** to **RQ 14**, as specified in Section 1.2.1. In particular, we formalize the layout of the blended aggregated search page and propose a utility-effort evaluation framework to capture the user behavior in order to answer the following questions:

> **RQ 11**: Do users agree with each other when assessing the preference of aggregated search pairs?
>
> **RQ 12**: Can we evaluate aggregated search pages (the whole aggregated search systems) that capture both effort and utility (relevance) in a formal way? How can we utilize (combine) both vertical relevance and document relevance when evaluating aggregated search pages?

[1] R and N represent a Relevant or Non-relevant result respectively.


Images for yoga poses
R
R
N
R
R
N
Yoga Poses | Basic to Advanced Yoga Poses | Asanas
Yoga Postures Step-By-Step - Index to Poses
Yoga Poses - MapQuest
R
R
R
N
Yoga Poses for Beginners: How-to, Tips, Benefits, Images, Videos
View topic - Yoga Poses Anxiety Panic Help
News for yoga poses
N
R
R
N
Videos for yoga poses
R
R
R
Kundalini Yoga For Weight Loss ...
4 Yoga Poses to Reduce Belly Fat ...
How to Do the Pigeon Yoga Pose ...

Page A
(Image+Web)

Page B
(Web)

Page C
(News+Web+Video)

Figure 7.1: Three Examplar Aggregated Search Pages for the Query "yoga poses" to Demonstrate the Complexity of Determining Aggregated Search Preference.

RQ 13: Do those aggregated search metrics possess strong *predictive power*, i.e. aligning with the real user preference of aggregated search pages?

RQ 14: Can we personalize the evaluation based on each types of user?

We address these questions by proposing a general framework for instantiating metrics that can evaluate the quality of aggregated search pages in terms of both *reward* and *effort* in a formal way. Specifically, we develop an approach that uses both topical-relevance and vertical-orientation information to derive the utility of any given aggregated search page. Our approach is flexible and takes into account any combination of items retrieved, any combination of verticals selected, and the positions of those results on the presented page.

We outline a novel approach for simulating aggregated search pages and collect a large set of user preferences over page pairs. We evaluate our approach by the annotated user preferences over a set of aggregated search pages for 56 topics and 12 verticals. Then we empirically demonstrate the fidelity of metrics instantiated from our proposed framework by showing that they strongly agree with the annotated user preferences of pairs of simulated aggregated pages. Furthermore, we show that our metrics agree with the majority user preference more often than the current diversity-based information retrieval metrics. Finally, we demonstrate the flexibility of our framework by showing that personalised historical preference data can improve the performance of our proposed metrics.

The remainder of this chapter is organized as follows. Related work is reviewed in Section 7.1. In Section 7.2, we formally outline the problem of aggregated search evaluation and list the assumptions made in this work. In Section 7.3, we propose a general framework from which we derive a number of metrics. A method for collecting the page preferences of users is outlined in Section 7.4. Subsequently in Section 7.5, these are used to evaluate the performance of our metrics against baseline ones. We also show how the performance of our metrics can be improved using training data. Conclusions are discussed in Section 7.6.

7.1 Related Work

Various works evaluating one component of an aggregated search system in isolation exist. Vertical selection in aggregated search has been studied in [Arguello et al., 2009, 2010, Li et al., 2008, Zhou et al., 2012a]. Much of this research aims to measure the quality of the set of selected verticals, compared with an annotated set obtained by collecting manual labels from assessors [Arguello et al., 2009, 2010, Zhou et al., 2012a] or derived from user interaction data [Li et al., 2008]. The annotation can be binary [Arguello et al., 2009, 2010] or graded [Zhou et al., 2012a]. The quality of a particular vertical selection approach is mostly evaluated with standard measures of precision and recall using the binary annotated set. Our work also evaluates this key component by utilising graded vertical-orientation information derived from a multi-assessor preferred vertical annotation set [Zhou et al., 2012a], as this allows for a more refined evaluation scheme.

Recent attempts to evaluate the utility of the whole aggregated search page [Arguello et al., 2011b, Ponnuswami et al., 2011b] consider the three key components of aggregated search (**VS**, **IS**, **RP**) together. Our work takes a similar holistic approach and proposes a general evaluation framework for measuring aggregated search page quality. For example, [Ponnuswami et al., 2011b] evaluate the utility of a page based on a user engagement metric (CTR). This evaluation framework requires large-scale user interaction data, which may not always be available. In addition, it is not feasible to collect user interaction data for all possible page combinations. Others [Bailey et al., 2010] evaluate the utility of the page by asking annotators to make assessments based on a number of criteria (e.g. relevance, diversity). Although this work is a comprehensive way to evaluate aggregated pages, it remains costly to gather assessments for all possible aggregated pages.

The most similar work [Arguello et al., 2011b] to ours collects preferences on block pairs from users and measures the page quality by calculating the distance between the page in question and the ideal (reference) page; the shorter the distance, the better the page. One advantage is that any possible combination of vertical blocks that form an aggregated page can be tested, from a block-oriented point of view (without regard to item selection). However, when the results retrieved for a vertical (block) change, the assessments previously gathered may not be reusable, as the preference will undoubtedly change accordingly. As our work follows the Cranfield paradigm, once the assessments (both item topical-relevance and vertical-orientation) are gathered, it can be applied to evaluate any possible aggregated search page (any combination of vertical selection, item selection and result presentation). Therefore, our work leads to a more robust, inexpensive, and reusable approach for evaluating aggregated search pages.

Topical diversity is an important topic. Various diversity-aware IR metrics have been proposed [Chapelle et al., 2009, Clarke et al., 2008, Sakai and Song, 2011], capturing the importance of each subtopic, the degree to which an item represents the subtopic, and the topical-relevance of the item. Diversity-based metrics can promote returned sets that are both topically relevant and diverse. A simplistic way of adapting these metrics to aggregated search is to treat subtopics as verticals and subtopic importance as vertical-orientation. In this way, all existing diversity-based IR metrics can be adapted to evaluate aggregated search. Although in principle suitable to evaluate aggregated search, diversity-based metrics are not appropriate for use with block-based pages where user behaviour is different; for instance user browsing behavior within a block containing images may be different to that within a block containing "general web" results. Furthermore, the various types of items (text, image, etc.) that need to be accounted for in an aggregated search scenario are not explicitly modelled in diversity-based metrics. For example, the effort in reading a piece of text is greater than the effort in viewing a picture. Our framework is better adapted to the task of aggregated search, and models all key components simultaneously.

Others [Santos et al., 2011] have proposed an aggregated search metric that captures both vertical diversity and topical diversity. It can be noted that the framework developed in this chapter can also be extended to incorporate topical diversity, but due to space limitations, we will leave this as future work.

7.2 Problem Formulation

We introduce some formal notation and outline some of the main assumptions used throughout this work.

7.2.1 Aggregated Page Composition

An aggregated search page P is composed of a set of blocks $\{B_1, B_2, ...B_n\}$, where each block B_i consists of a set of items $\{I_{i1}, I_{i2}, ...I_{im}\}$. An item can be a "general web" page or a vertical result. Only snippets of each item appear on the aggregated search page. We make several assumptions[2] about the page P: (i) results are presented into blocks from top to bottom, and within each block, items are shown either from left to right (Image, Video) or from top to bottom (News, Recipe); (ii) each block B_i consists of items originating from only one vertical; (iii) only one block of each type is placed on a page (with the exception of "general web" blocks); and (iv) a block consists of one 'general web' item or k vertical items. This is different to previous work [Arguello et al., 2011b, Ponnuswami et al., 2011b] where vertical block could be embedded into only three positions of "general web" results (top of the page, middle of the page, bottom of the page). We relax this assumption and allow a vertical block to be slotted between *any* two "general web" blocks on the page.

7.2.2 Relevance and Orientation

Our objective is to develop metrics that measure the quality of any possible aggregated search page. The metrics must work regardless of the selected verticals, the items retrieved from each vertical, and where the vertical results are positioned on the page. To achieve this, we assume that the following two types of relevance assessments are available:

- The topical-relevance of each item, which is an assessment indicating whether a given item I_{ij} within block B_i is topically relevant to a topic q[3]. This is denoted

[2] These assumptions are made in accordance with existing aggregated search systems.

[3] In this work, we assume that topical-relevance assessments are binary.

$qrel(I_{ij}|q)$.

- The user specific vertical-orientation [Zhou et al., 2012a], which is a value between zero and one indicating the fraction of users that prefer a page to contain items from the vertical V_i rather than "general web" results for a topic q. This is denoted $orient(V_i|W, q)$.

The two relevance assessments are assumed to be made independently. The concept of vertical-orientation [Zhou et al., 2012a] reflects the perceived usefulness of a vertical from the user perspective prior to viewing vertical results and without regard to the quality of the vertical results. The vertical-orientation assessment is obtained by comparing each vertical in turn to the reference "general web" vertical, by asking users whether items from this vertical are likely to improve the quality of a standard web page. Consequently, the vertical orientation of the Web ($orient(W|W, q)$) is deemed to be 0.5, as we can imagine that a user would randomly select a page when presented with two similar "general web" pages. The topical-relevance assessment of each item contributes to the measurement of relevance for each retrieved result. This type of assessment can be made using similar pooling techniques [Jones and van Rijsbergen, 1975] to those used in TREC.

With these two assessment types, we assume that a user obtain the highest reward by reading the most topically relevant item, originating from the most highly oriented vertical, first. With this assumption, only the vertical (or verticals) with a higher orientation than the "general web" ($orient(V|W, q) > 0.5$) should be presented on the aggregated search page; all other verticals should be suppressed.

7.2.3 User Interaction Behaviour

We make some assumptions about how users interact with an aggregated search page P:

- The user examines each page one block at a time. When the user reads page P, a block B_i on the page P has a certain probability of being examined. This probability denoted $Exam(B_i)$ is estimated depending on the type of browsing model assumed.
- After the user decides to examine a block B_i, we assume a static user browsing behavior within the block; the user reads all the items I_{i1} to I_{im} within that block.

Given that our metrics are based on average user and that there is usually only a limited number of items per block, this simple within block user browsing model is appropriate.

7.3 Evaluation Framework

We aim to develop metrics that evaluate an aggregated search page similarly to how a user might. Given two pages P_1 and P_2, we wish to measure their effectiveness in satisfying a user information need using a utility function $Util(P)$. If a user prefers P_1 over P_2 for a given query, the utility measure should lead to $Util(P_1) > Util(P_2)$.

Following [Dupret, 2009], the utility of a page is determined by *reward* and *effort*. A page with a high utility should satisfy the average user information need with relatively little effort. We define the utility metric $Util(P)$ of the page P based on all blocks $\{B_1, B_2, ...B_n\}$ on the page. When a user reads page P, a block B_i on the page P has a certain probability $Exam(B_i)$ of being examined. This probability might depend on the position of the block presented, the snippet type of the items (image, text) within the block, or the satisfaction level after reading previous blocks B_1 to B_{i-1}. The probability $Exam(B_i)$ can be estimated depending on the type of browsing model assumed.

After the user decides to read block B_i, he/she will be rewarded with some gain $G(B_i)$ coming from reading all the items I_{i1} to I_{im} within that block. Here we assume that the topical-relevance of the item snippet is a good indication of the relevance of the item itself. Therefore, by reading all the items within the block B_i, the user will also have spent some effort $E(B_i)$ in reading this block. Therefore, based on our assumptions, we define the utility of the page $Util(P)$ as the expected gain of reading a page divided by the expected effort spent:

$$Util(P) = \frac{\sum_{i=1}^{|P|} Exam(B_i) \cdot G(B_i)}{\sum_{j=1}^{|P|} Exam(B_j) \cdot E(B_j)} \tag{7.1}$$

where $|P|$ is the number of blocks on page P. To ensure suitable normalisation over a set of queries, we define a normalized utility score $nUtil(P)$, similar to $nDCG$ [Järvelin and Kekäläinen, 2002]. We normalise the score of the utility of page P by that of the ideal page P_{ideal}:

$$nUtil(P) = \frac{Util(P)}{Util(P_{ideal})} \tag{7.2}$$

Until now, we have defined a general evaluation framework for any aggregated search page that considers both reward and effort simultaneously. Consequently, for two pages P_1 and P_2, we can say P_1 is better than the other when $nUtil(P_1) > nUtil(P_2)$. In the following sections, we instantiate the gain $G(B_i)$, the effort $E(B_i)$, and the examination probability $Exam(B_i)$ of the blocks. We then outline how to normalise the $Util(P)$ metrics by constructing an ideal page. Finally, we incorporate a simple personalisation parameter that captures the degree to which a user prefers vertical diversity on an aggregate search page.

7.3.1 Gain of Reading a Block

Given a block B_i containing a set of items (I_{i1}, I_{i2}, ... I_{im}) originating from vertical V_j, we would expect that if the vertical is highly oriented given the query, the user will achieve a higher gain. We denote this block orientation as *Orient*, which is related to the task of *vertical selection*. Furthermore, we would expect that the more topically relevant items a block contains, the higher the gain for the user. We denote the topical-relevance of the block as *Topic*. Before combining these two factors, we define the gain relating to the vertical-orientation of the block B_i:

$$Orient(B_i, \alpha) = g(orient(V_j|W, q), \alpha) \tag{7.3}$$

where $orient(V_j|W, q)$ is a value between 0 and 1. The function $g()$ is used so that the relative gain of the vertical can be altered using a tuning parameter α. The $orient(V_j|W, q)$ value is defined as the fraction of users that would prefer the vertical V_j to be added to the "general web" results W. As the "general web" is the pivot to which verticals are added, if $orient(V_j|W, q) > 0.5$, then adding the vertical should be rewarded. If $orient(V_j|W, q) < 0.5$, the gain of the block should be less than the "general web" results (i.e. 0.5). Therefore, we use a pivot at the 0.5 value through which $g()$ must pass. The following function satisfies these criteria:

$$g(x, \alpha) = \frac{1}{1 + \alpha^{-log_{10}(x/(1-x))}} \tag{7.4}$$

A graph of the function $g(x, \alpha)$ is shown in Figure 7.2. This function controls how much the gain increases as the vertical-orientation level increases. When α is small ($1 < \alpha < 10$), we obtain a more steep curve; highly oriented verticals are more rewarded, and conversely, low orientated verticals are more penalised. When α equals to 10, the reward is exactly the same as the vertical orientation $orient(V_j|W, q)$.

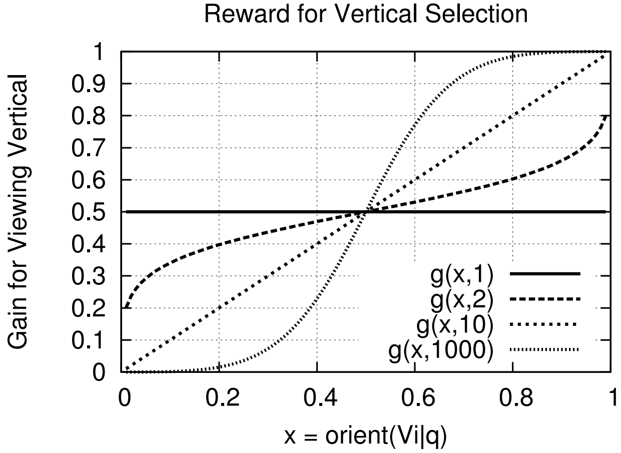


Figure 7.2: Function $g()$ for Controlling Reward on Orientation with Various Parameter α.

Now we define the gain relating to topical-relevance of the items within B_i:

$$Topic(B_i) = \sum_{k=1}^{|B_i|} qrel(I_{ik}|q) \tag{7.5}$$

$|B_i|$ is the number of items within block B_i and $qrel(I_{ik}|q)$ is the binary relevance assessment of the item I_{ik}. In short, we use the sum of the binary relevance judgments of the items as the topical-relevance gain of all the items within the block B_i.

Now that we have defined the gain of a block in terms of both vertical-orientation and topical-relevance, we combine these in a suitable manner. Specifically, we combine the

Table 7.1: Effort of Reading each Category.

Snippet Category	“image”	“text”	“video”
Effort	1	3	6

gain based on the above two criteria:

$$G(B_i) = Orient(B_i, \alpha) \cdot Topic(B_i) \tag{7.6}$$

where α is the tuning parameter as described above. We combine these two factors in an independent manner as both vertical-orientation and topical-relevance are related to the quality of the block. Either a low oriented block (low $Orient(Bi)$) or a topically irrelevant item (low $Topic(Bi)$) would result in an unsatisfied user.

7.3.2 Effort of Reading a Block

We now consider the effort $E(B_i)$ spent in examining a block B_i. Based on the assumed block-based user browsing behavior, the effort of examining a block is defined as the accumulative effort of reading all the items within it:

$$E(B_i) = \sum_{k=1}^{|B_i|} E(I_{ik}) \tag{7.7}$$

where $|B_i|$ is the number of items within block B_i, $E(I_{ik})$ is the effort spent in reading the item I_{ik}.

Several factors may affect the effort spent in examining an item $E(I_{ik})$: the media type of the snippet (text, image) or the size of the snippet (text length). We assume that there are only three categories of item snippet (“image”, “text” and “video”). Furthermore, we assume that “image”, “text” and “video” have a standard size. Based on [Xue et al., 2008], the time taken to assess the relevance of an image is estimated 2.34 seconds, while the time taken to assess a text snippet is 7.02 seconds. We extrapolate that a video takes twice as much time to assess as a text[4] (14 seconds). Therefore, the relative effort taken to examine each snippet type is shown in Table 7.1 and is used as the unit of effort. These settings are not optimal and have been chosen heuristically after a review of the literature. Identifying more optimal settings is outside the scope of this work.

7.3.3 Examination Probability for a Block

We concentrate on defining the user browsing model for examining a block $Exam(B_i)$ on a page. Several models exist [Chapelle et al., 2009, Chen et al., 2012, Moffat and Zobel, 2008] that aim to predict the probability with which a user will examine an item. Position models [Moffat and Zobel, 2008] use only the position of the item in a result

[4] We assume that users need to open and view the video item to assess its topical-relevance.

set. The cascade model [Chapelle et al., 2009] uses the relevance of the items previously examined, the intuition being that a sufficiently satisfied user will not continue to examine extra items. Motivated by the fact that users tend to be attracted by vertical results and the visual attention on them will increase the examination probability of other nearby web results, the attention model [Chen et al., 2012] aims to capture the visual attractiveness of the page. We do not propose a new user browsing model for aggregated search. Rather, we adopt these different models and incorporate them into our framework, namely the position examination models **DCG** [Järvelin and Kekäläinen, 2002] and **RBP** [Moffat and Zobel, 2008], the cascade model **ERR** [Chapelle et al., 2009], and the attention model **ATT** [Chen et al., 2012].

To adapt **ERR** to block examination, we assume that the satisfaction of viewing previous blocks is defined as the average gain of viewing each item within the block. For **ATT**, β_{dist} is the distance between the item under consideration and the closest vertical that has the attention bias (image and video). As we do not have access to query logs to accurately estimate the attention bias parameter ζ, instead of assuming that ζ is a position-specific parameter, we assume that ζ is a global variable that is constant for all positions. In addition, there will be attention-bias only when results from image or video verticals are presented on the page. The standard ζ is obtained by exploring the optimal setting in a development set.

7.3.4 Normalisation Using the Ideal Page

Table 7.2: Summarisation of Utility Metrics for Aggregated Search.

Metric	Examination Model $E(k)$	Parameter	Utility
AS_{DCG}	$\frac{1}{log(k+1)}$	α	
AS_{RBP}	β^{k-1}	α, β	$Util(P)$
AS_{ERR}	$\frac{\prod_{j=1}^{k-1}(1-\frac{G(B_j)}{\|B_j\|})}{k}$	α	$= \frac{\sum_{i=1}^{\|P\|} G(B_i)\cdot E(i)}{\sum_{j=1}^{\|P\|} E(B_j)\cdot E(j)}$
AS_{ATT}	$[(1-\frac{1}{log(k+1)})\cdot\beta_{dist}+\frac{1}{log(k+1)}]\cdot\zeta$	α, ζ	

A summary of the non-normalised utility metrics that can be instantiated in our framework are listed in Table 7.2. We have a suite of metrics that reward pages that contain highly oriented verticals, contain topically-relevant items, promote topically-relevant blocks earlier on the page, for less effort. The utility metrics must be normalised by the ideal aggregated page. To obtain the latter, we require a brute-force approach that calculates the metric score for all pages, and then selects the page with the maximal score as the ideal page ($arg\_max(Util(P))\forall P$). This approach is not viable, given the number of possible combinations of various components of aggregated search. Therefore, we use a greedy algorithm to select a subset of aggregated pages from all the pages that exist, and only select the optimal page from this set. The idea is to use a simple metric for each component, and only select the pages that perform optimally for all those components. This is described in Section 7.4.2, where the simulation of aggregated page pairs is discussed.

7.3.5 Personalised Utility Metrics

Previous research [Zhou et al., 2012a] has shown that different users have different preferences with regard to the type of vertical. A vertical with low orientation to a query for the average user may still be beneficial to users that prefer a very diverse information space. Therefore, we define a personalised vertical diversity preference factor to capture this scenario. We achieve this by linearly combining the normalised utility of the page with the vertical recall. This introduces a personalised preference parameter λ_i:

$$I\_Util(P, \lambda_i) = (1 - \lambda_i) \cdot nUtil(P) + \lambda_i \cdot vRecall(P) \tag{7.8}$$

where λ_i is a parameter between 0 and 1 for user i, and controls the trade-off between vertical diversity and the quality of the aggregated search page. $vRecall(P)$ represents the fraction of all verticals that are presented on page P. The larger λ_i is, the more the user prefers a page with items originating from different verticals (high vertical diversity).

7.4 Collecting Pairwise Preference Assessments

To validate the fidelity of our metrics (how they agree with actual user preferences of aggregated search pages), we collected a set of pairwise preference assessments over aggregated page pairs. We first present the data and material used for this purpose. We then simulate a set of aggregated search pages that vary in different levels of quality for each topic. Afterwards, we select a set of page pairs (two simulated pages) for each topic. Finally, we collect preference assessments for the page pairs for all topics. We outline some statistics and analysis of the assessments gathered.

7.4.1 Data

We use an aggregated search test collection [Zhou et al., 2011] created by reusing the existing web collection ClueWeb09 B, as described in Chapter 4. This test collection consists of a number of verticals (listed in Table 4.1), each populated by items of that vertical type, a set of topics (320) expressing information needs relating to one or more verticals, and assessments indicating the topical-relevance of the items and the perceived user-oriented usefulness of their associated verticals to each of the topics. The verticals are created either by classifying items in the web collections into different genres (e.g. Blog, News) or by adding items from other multimedia collection (e.g. Image, Video). The topics and topical-relevance assessments of items that vary in genres are obtained by reusing assessments developed in TREC evaluation tasks (TREC Web Track and MillionQuery Track). The vertical-orientation information of each topic [Zhou et al., 2012a] is obtained by only providing the vertical names (with a description of their characteristics) and asking a set of assessors to make pairwise preference assessments, comparing each vertical in turn to the reference "general web" vertical ("is adding results from this vertical likely to improve the quality of the ten blue links?"). This procedure follows Section 5.2.

We select a subset of topics from which to collect assessments. We ensure that this subset of topics still conforms to the real-world distribution of aggregated search covering a wide range of needs with different highly oriented verticals. Therefore, we selected 56 topics detailed in Table 7.3.

Table 7.3: Distribution of Number of Selected Topics Assigned to Various Highly Oriented Verticals (in Total 56 Topics).

Verticals	Image	Video	Recipe	News	Books	Blog
Topic Num	4	3	3	4	3	3
Verticals	Answer	Shopping	Discuss	Scholar	Wiki	Web-only
Topic Num	3	3	5	3	12	10

7.4.2 Simulating Aggregated Search Pages

For each topic, we simulate a set of aggregated search pages. As indicated in Section 3, we assume that a page consists of ten "general web" blocks (one "general web" page is a block) and up to three vertical blocks dispersed throughout those ten blocks (where each vertical block consists of a fixed number of three items). Recall that there are three key components of an aggregated search system that can be varied: (i) Vertical Selection (**VS**); (ii) Item Selection (**IS**); and (iii) Result Presentation (**RP**). We generate pages by simulating an aggregated search system in which the three components vary in quality.

The assessments for vertical-orientation were created by gathering annotations across several users. For the process of varying **VS**, for a given vertical V_i and query q, we consider the vertical to have a high vertical orientation if $orient(V_i|W,q)$ is greater than 0.75 [5]. We simulate four different vertical selection strategies, namely *Perfect*, *ReDDE*, *CORI*, *Bad*. *Perfect* selects all the highly oriented verticals, while *Bad* randomly selects the maximum number (three) of lowly oriented verticals. *ReDDE* and *CORI* rank the verticals according to the ReDDE [Si and Callan, 2003a] and CORI [Callan et al., 1995] resource selection approaches, and select the top K ranked verticals.

For **IS** we simulate three potentially different levels of relevance. These are *Perfect*, *BM25*, and *TF*. *Perfect* selects all items in the vertical that are topically relevant. *BM25* and *TF* select the top three ranked items from the rankings provided by the BM25 and a simple TF (term frequency) weighting respectively, with the PageRank score as a prior for both *BM25* and *TF*.

For **RP**, we simulate three different result presentation approaches: *Perfect*, *Random* and *Bad*. *Perfect* places the vertical blocks on the page so that gain could potentially be maximised, i.e. all the relevant items are placed before non-relevant items. However, if

[5] We select the threshold as 0.75 as 75% assessors majority preference is a suitable percentage whereby the assessments are neither too noisy (50%) or stringent (100%). Furthermore, it creates a vertical intent distribution across the topics that realistically conforms to the real-world [Arguello et al., 2009].

these items are part of a vertical, we position the highest orientated vertical first. *Random* randomly disperses the vertical blocks on the page while maintaining the position of the "general web" blocks. *Bad* reverses the perfectly presented page.

By varying the quality of each of the three key components, we can vary the quality of the result pages created by an aggregated search system in a more controlled way. For each topic, we can create 36 (4 × 3 × 3) pages[6]. In addition, the snippet of each item is automatically generated by the Lemur Toolkit and the presentation style conforms with typical search page presentation (presenting the vertical name in front of vertical results). Using this approach we can create a near ideal aggregated page for a query by using *Perfect* **VS**, *Perfect* **IS**, and *Perfect* **RP**. This is a greedy approach to the problem and is used as our method of normalisation for $nUtil$.

7.4.3 Constructing and Selecting Page Pairs

We now describe the selection of page pairs so that they can be presented to a user for judgment. One way to achieve this is to randomly sample two aggregated search pages, and collect a sufficient set of user preference judgments. However, following [Arguello et al., 2011b], we attempt a broad categorisation of the aggregated search pages into "bins" according to page quality, i.e. H (High), M (Middle) and L (Low). We can then provide a more in depth analysis of the performance of the metrics over different regions of the page space.

Although we do not know the quality of all the pages, we can roughly estimate the page quality using the quality of the components that created the page. We estimate this by assuming that the three components contribute equal importance to the quality of the page. We then evaluate each component respectively using a suitable metric. The quality score of the page is determined by linearly combining the metric score for each component. This is a coarse approach of determining the quality of the page. We use the F-measure (VS), Mean Precision (IS), and Kendall-tau correlation (RP). We then rank all the pages according to the three linearly combined metrics and evenly categorise the pages in the ranking into "H", "M" and "L" bin respectively.

We now have a method of comprehensively analysing how various metrics perform over the whole page space by selecting pages from these pre-assigned bins. Specifically, we have six bin pairs, H-H, H-M, H-L, M-M, M-L, L-L, which uniformly represent all the entire page space for the queries (albeit in coarse intervals). For each pair of bins, we randomly select 8 page pairs from it. Consequently, we select in total 48 (6 × 8) page pairs for each topic.

7.4.4 Collecting Pairwise Preference Assessments

Our preference assessment data is collected over the Amazon Mechanical Turk crowdsourcing platform, where each worker was compensated $0.01 for each assessment made. A page pair was presented with the topic (title and description) shown in the upper position of the assessment page. This was followed by a pair of aggregated pages shown side-by-side. The assessor was provided with three options when making the assessments: "left page is better", "right page is better" and "both are bad". The latter option

[6] Certain combinations of VS, IS, and RP do not create unique simulated pages.

captures the scenario where a user is confused due to the poor page quality[7]. For each page pair, we collect four assessments (from four different assessors). The total number of assessments made during this preference collection process was 10752 ($56 \times 48 \times 4$). Following [Sanderson et al., 2010], a quality control was ensured by including 500 "trap" HITs. Each "trap" HIT consists of a triplet (q, i, j) where either page i or j was taken from a query other than q. We interpreted an assessor preferring the set of extraneous results as evidence of malicious or careless assessments and assessors who failed more than two trap HITs were discarded.

7.4.5 Analysis of Assessments

Of the 203 assessors who contributed HITs, 39 had their assessments removed from the assessment pool due to failing more than 2 trap HITs. For the remaining 164/203, participation followed a power law distribution where about 12% (20/164) of the assessors completed about 60% (6522/10752) of our HITs. We also found out that assessors rarely select the "both are bad" options provided as only 7% (684/10752) of the assessments are of this option.

We first want to answer the following question:

> **RQ 11**: Do users agree with each other when assessing the preference of aggregated search pairs?

Therefore, we measured annotator agreement of preferences of aggregated page pairs using Fleiss' Kappa [Fleiss, 1971] (denoted by K_F), which corrects for agreement due to chance. Fleiss' Kappa is convenient because it ignores the identity of the assessor-pair, and is designed to measure agreement over instances labeled by different (even disjoint) sets of assessors. The results are shown in Table 7.4.

We observe that assessor agreement on presentation-pairs was $K_F = 0.241$, which is considered *fair* agreement [Fleiss, 1971]. This result is similar to previous research [Arguello et al., 2011b, Zhou et al., 2012a], which re-affirm that evaluating aggregated

[7] The option "Both are good" is not included because this information can be potentially obtained by investigating inter-assessor agreement for definite preferences.

Table 7.4: Statistics of User Preference Assessment Agreement over Various Quality Bins.

bins	4/4	3/4	Kappa agreement
all	2231	8051	0.241
H-H	347	1396	0.238
H-M	427	1354	0.283
H-L	461	1318	0.317
M-M	287	1424	0.192
M-L	394	1327	0.261
L-L	315	1332	0.210

search is not an easy task, and that various users have their own assumptions about what a good page is. Of all 10752 aggregated page-pairs, 8051 (74.8%) had a majority preference of at least 3/4 and only 2231 (20.7%) had a perfect 4/4 majority preference. It is perhaps not surprising that assessor agreement is not high as agreement on page-pairs requires that assessors make similar assumptions about the cost of different types of errors. Furthermore, the low inter-assessor agreement may be explained by the fact that users make different assumptions regarding the importance of each aggregated search component (**VS**, **IS**, **RP**). Alternatively, it may be that assessors have a hard time distinguishing between good presentations. Following previous research [Arguello et al., 2011b], given this low level of inter-assessor agreement, rather than focusing on the metrics agreement with each individual preference, we focus on their agreement with the majority preference (3/4 or greater, and 4/4) in the evaluation.

7.5 Evaluation

We investigate the fidelity[8] [Voorhees, 2003] (also as known as *predictive power* [Sanderson et al., 2010]) of the proposed metrics. We leave an investigation on the reliability of the metric (discriminative power [Sakai and Song, 2011]) for the next chapter. We aim to answer the following questions:

> **RQ 12**: Can we evaluate aggregated search pages (the whole aggregated search systems) that capture both effort and utility (relevance) in a formal way? How can we utilize (combine) both vertical relevance and document relevance when evaluating aggregated search pages?
>
> **RQ 13**: Do those aggregated search metrics possess strong *predictive power*, i.e. aligning with the real user preference of aggregated search pages?

To demonstrate the fidelity of our four metrics (AS_{DCG}, AS_{RBP}, AS_{ERR} and AS_{ATT}), we compare them with existing IR metrics. We utilise both user-oriented IR metrics capturing topical-relevance ($nDCG$ [Järvelin and Kekäläinen, 2002], $P@10$), and diversity-aware metrics (α-$nDCG$ [Clarke et al., 2008], D-$nDCG$ [Sakai and Song, 2011], $D\#$-$nDCG$ [Sakai and Song, 2011], IA-$nDCG$ [Agrawal et al., 2009]) which we adapt to incorporate vertical diversity. We select the latter as they are the most prevalent user-oriented IR metrics. Their adaptation is as follows: (i) we replace subtopic importance with $orient(V|W, q)$; (ii) we substitute the user model for ranks to the one that applies to blocks; and finally (iii) we normalise according to the ideal aggregated search page.

To measure the performance of the metrics, we calculate the percentage of agreement (percentage of those pairs for which the metric agrees with the majority preference of users). The larger the percentage of agreement, the more accurately the metric can predict the user preference of any aggregated search page pairs, and the higher the metric fidelity. A two-tailed t-test (significant at the $p < 0.01$ level, denoted by ▲ or ▼) is used to show which metric correlates more significantly with the user preferences[9].

[8] The extent to which an evaluation metric measures what it is intended to measure.

[9] We also used the sign test [Arguello et al., 2011b]. For all page pairs with majority of preference, our proposed metrics performed significantly better than random. Since we are interested in comparing metrics,

7.5.1 Standard Parameter Settings

To answer **RQ 12** and **RQ 13**, we carry out a set of experiments where we employ the prevalent standard parameter settings for the metrics used in IR experiments. We utilise the standard log discount function for all DCG related metrics (AS_{DCG}, $nDCG$, α-$nDCG$, D-$nDCG$ and D#-$nDCG$). We set the α parameter in α-$nDCG$ to 0.5 and γ to 0.5 for D#-$nDCG$. For our proposed metrics, we set $\alpha = 10$ (a linearly increasing vertical-orientation function) and $\lambda_i = 0.0$ (no personalised vertical diversity preference) as the standard parameters. For the user persistence parameter in AS_{RBP}, we set $\beta = 0.8$ as this value best correlates with the user browsing behavior from a real-world query-log data [Moffat and Zobel, 2008]. These standard settings instantiate a simple metric (e.g. AS_{DCG}) similar to existing topical diversity-aware metrics that incorporate subtopic importance probability (D-$nDCG$). The standard ζ of AS_{ATT} is obtained by exploring the optimal setting in a development set that contained 500 preference page pairs that contain visually attractive results (results coming from Image and Video).

Our evaluation, the fidelity of the metrics, thus focuses on the agreement (of each metric) with the user preferences over the set of aggregated search results. As we have already categorised page pairs into various quality "bins" (H-H, H-M, H-L, M-M, M-L, L-L), we report the experimental results over different bin pairs, in order to understand each metric performance over the whole evaluation space. Our experiments have two parts: (i) when fixing the assumed user browsing model (e.g. DCG), we compare the performance of our proposed metrics with existing IR metrics; (ii) under the proposed framework, we compare user models to investigate which ones make more accurate prediction of the user preferences on aggregated page pairs.

Comparison of Metrics

We present results for a majority preference of 3/4 or greater, or 4/4, in Table 7.5. The significance is calculated in comparison with one of the proposed metrics, AS_{DCG}. Our metrics have higher agreement with user preferences for the H-M, H-L and M-L bins compared to the less discriminative bins (H-H, M-M, or L-L). In addition, for page pairs with higher majority user agreement (4/4 instead of 3/4), our metrics tend to make more accurate prediction of the user preferences. After closer examination, we observe that the metrics agreement with the majority user preference is higher on pairs where there is greater consensus between assessors. This is similar to reported in [Arguello et al., 2011b].

We also observe that overall the proposed aggregated search metrics (AS_{DCG}) work better than existing IR metrics ($nDCG$ and P@10). They have a significantly better performance across almost the entire metric space. This is not surprising given that the proposed metrics incorporate aspects unique to aggregated search (vertical-orientation), which can affect user preferences. Indeed, when the page quality is expected to be high, traditional IR metrics that do not consider vertical-orientation perform worse than the proposed metrics. But it is worth noting that $nDCG$ performs significantly better than other metrics on L-L page pairs. This might be because as the returned verticals are of

we do not report the sign test outcomes.

low orientation, and for these types of page pairs, simply measuring topical relevance of items might correlate more with the user browsing behavior than considering the additional vertical orientation; when assessing two low-quality pages, the user is trying to find more topically relevant items, without regard to the orientation of the vertical.

For the diversity-aware metric, α-$nDCG$ performs significantly worse than the proposed metrics. This is because α-$nDCG$ implicitly penalises the within vertical redundancy of items. This evaluation strategy is not appropriate when presenting results from the same vertical in a block. A close examination shows that this degraded performance is due to the over-penalisation for items within each vertical. Although recent research [Leelanupab et al., 2011] has suggested that α may be tuned on a per query basis to either promote or discount extra items from the same sub-topic (vertical), we leave this for future work. In addition, instead of fully utilising the graded $orient(V|W,q)$ information, α-$nDCG$ treats relevant verticals in a binary sense, another reason that may cause the degraded performance.

The other existing diversity-aware metric D-$nDCG$ performs comparably well. This is not surprising as when employed with standard parameter setting, D-$nDCG$ is most similar to the proposed aggregated search metrics (AS_{DCG}). The major difference is that AS_{DCG} captures the effort of examining result snippets of different types. D#-$nDCG$ performs significantly worse than D-$nDCG$ over the entire simulated page space used for evaluation in the context of aggregated search. This proves that simply promoting vertical diversity without considering vertical-orientation can degrade the evaluation performance. In addition, as we will see later, because of the various users vertical diversity preference, personalised vertical diversity can be a better strategy for the evaluation of aggregated search. Finally, IA-$nDCG$ also performs considerably worse than AS_{DCG}. A close examination suggests that this is due to the over-rewarding of the vertical results in a page.

When we assume a uniform effort distribution of the resulting snippets, which can be of various types, the metric performances decrease from 67.3% to 65.6%. However, this decrease is not statistically significant. This might be due to the small number of topics promoting image or video vertical results. Estimation of the efforts associated with reading snippet of various types on a large-scale dataset is needed.

Comparison of User Models

For the proposed metrics with various user models (AS_{DCG}, AS_{RBP}, AS_{ERR} and AS_{ATT}), their agreements with the users majority preference (3/4 or greater) are shown in Table 7.6[10]. We observe that the metric agreements are comparatively similar; although, overall, the metrics based on position-based user models (AS_{DCG} and AS_{RBP}) perform consistently better than the adapted cascade model metric AS_{ERR} or the attention-based model AS_{ATT}.

[10] The results of metric agreement with 4/4 users majority preference is similar and is, therefore, not included due to space limitations.

Table 7.5: Metric Agreements with Various User's Majority Preference: Proposed Metric vs. Baseline Metrics.

majority preference	bins	AS_{DCG}	D-$nDCG$	D#-$nDCG$	IA-$nDCG$	α-$nDCG$	$nDCG$	P@10
3/4 or greater	all	67.3%	65.9%	62.9%▼	64.3%	62.4%▼	60.1%▼	53.9%▼
	H-H	61.4%	60.4%	57.2%▼	57.0%▼	54.0%▼	53.3%▼	49.5%▼
	H-M	74.3%	72.3%▼	68.8%	71.1%	60.5%▼	63.1%▼	61.2%▼
	H-L	78.0%	78.4%	76.3%	75.8%	73.3%▼	67.9%▼	58.3%▼
	M-M	64.7%	62.7%▼	64.2%	64.8%	64.9%▼	61.1%▼	51.2%▼
	M-L	72.4%	68.1%	67.1%▼	65.8%▼	70.2%▼	67.3%▼	55.1%▼
	L-L	51.3%	52.6%	53.2%	53.1%	51.7%	54.7%▲	47.3%▼
4/4	all	71.1%	69.4%	64.8%▼	67.7%	63.1%▼	60.9%▼	54.1%▼
	H-H	68.2%	65.4%	56.3%▼	62.1%▼	53.1%▼	52.4%▼	52.3%▼
	H-M	76.3%	76.0%	70.1%▼	78.2%	62.0%▼	64.8%▼	58.1%▼
	H-L	77.6%	78.9%	76.9%	78.3%	74.1%	65.9%▼	56.7%▼
	M-M	67.3%	65.1%	65.1%	63.7%	63.4%	62.5%	49.4%▼
	M-L	75.2%	72.4%	66.5%▼	68.4%▼	72.0%	68.3%▼	57.2%▼
	L-L	61.1%	57.8%▼	51.3%▼	54.5%▼	51.9%▼	52.6%▼	52.3%▼

Table 7.6: Proposed Metric Agreements with 3/4 User Majority Preferences: Comparison of User Examination Models.

bins/metrics	AS_{DCG}	AS_{RBP}	AS_{ERR}	AS_{ATT}
all	67.3%	67.7%	63.8%▼	66.9%
H-H	61.4%	62.1%	53.1%▼	60.5%
H-M	74.3%	75.4%	78.2%▲	72.1%▼
H-L	78.0%	80.3%	79.1%	77.4%▼
M-M	64.7%	65.2%	56.7%▼	66.3%
M-L	72.4%	71.9%	64.9%▼	70.0%
L-L	51.3%	48.8%▼	54.1%	54.5%

We further see that comparatively AS_{ERR} performs better on H-M and H-L bins and worse on others. The degraded performance might be due to the fact that only binary topical-relevance assessments (of items) are available and the metric largely rewards the top relevant results. This also partly explains why AS_{ERR} performs particularly well between high quality pages (highly oriented and relevant results are presented at the top of the page) and low quality pages. It is most likely that instead of considering the entire page, most assessors looked only at the early results of the page when assessing.

However, surprisingly, by incorporating attention bias (of visually attractive vertical results) into the position-based model, the performance of the metric AS_{ATT} degrades, compared with AS_{DCG}. This might be due to the inaccurate estimation of the attention bias ζ from our small-scale experiments. After closer examination, it may be that assessors have a considerable preference bias on pages that contain visually attractive results (image, video) [Arguello et al., 2011b]. Therefore, the preference assessment between pages containing image and video verticals may be noisier, which could result in a natural bias for those types. Further experiments are needed to explain and understand this bias and its effect.

In comparing AS_{DCG} and AS_{RBP}, although it is observed that AS_{RBP} performs slightly better for page pairs consisting of pages with high quality agreements, the result is not significant. As the only difference between AS_{DCG} and AS_{RBP} is the position-based discounting factor (the user browsing model), the slight improvement is caused by the different user model. This user browsing modelling factor is examined in more detail later.

Summary

Although the results of our proposed metrics are promising when compared with existing IR metrics, the results should be treated with caution as the agreement is not substantial (the best performance is 67.7% from our proposed metric AS_{RBP}). After a close examination of the user preferences, compared with the metric prediction, the reasons for this include: (i) the vertical-orientation annotations [Zhou et al., 2012a] may not fully agree with the real user preference of verticals (they are noisy estimations); and (ii) although three key components of aggregated search are captured, we have only used simple default values for some of the parameters. This motivates further experiments that aim to learn personalisation parameters from historical data.

7.5.2 Learning for Metrics

In this section, we aim to answer:

> **RQ 14**: Can we personalize the evaluation based on each types of user?

We can improve the performance of our metrics by learning suitable parameter settings using training data, thus addressing the research question **RQ 14**. We only use AS_{RBP} as an example. We recall that AS_{RBP} has three parameters: α that controls the degree to which vertical orientation is rewarded; β that controls the user browsing behavior in terms of user persistence; and λ_i that controls the degree to which a user prefers a diverse aggregated page.

Training is done in two stages. First, we learn suitable values for α and β independently of λ_i. We categorised the user preference data into five sets and use five-fold cross validation for training and testing. We set $\lambda_i = 0.0$ (users do not prefer vertical-based diverse results unless the vertical provides better results) and iterate through different settings of values: α (from 1 to 100) and β (from 0.5 to 1.0). The optimal combination is obtained with $\alpha = 7.0$ and $\beta = 0.85$ indicating that users generally favour results that contain highly-oriented verticals, and that users do not have a persistent browsing behaviour (they care more about the results returned in a high position in the page). The corresponding results are shown in Table 7.7. The performance of the metric is improved over the standard parameter settings from 67.7% to 72.6%. This improvement is due to the better estimation of two parameters α and β concerned with two main aspects of aggregated search, vertical selection and result presentation. By learning from historical data, AS_{RBP} (and other metrics) can better capture these two aspects. Second, we fix the optimal settings for α and β and learn personalised user preference parameters for diversity (λ_i). Although not optimal, this is sufficient to analyse the "personalisable" parameter independently of others.

As we need sufficient data for learning the parameter, we only test this over the top twenty "head" assessors who made most of the assessments. Like with previous setting, for each assessor, we separate assessor data into five sets and use five-fold cross validation to train and test. For the overall performance, we average the performance for all those assessors. The results are also shown in Table 7.7. The optimal setting for λ_i varies from 0.15 to 0.4 among different assessors whereas the average optimal setting for those assessors is 0.23. Similar to [Zhou et al., 2012a], this demonstrates that each user has his/her own understanding and preference over the diversity of the results. We can report that by using this personalised parameter, the prediction of the metric agreement with the majority of user preference is improved significantly, from 72.6% to 75.9%. This effectively illustrates that aggregated search can be improved if we have a better understanding of each user preference over the diversity of results. This is particularly useful for systems that can gather personalised interaction data for their users.

Table 7.7: Learned AS_{RBP} Metric Agreements with User's Majority Preferences for All Page Pairs.

Parameter	Standard	Optimal α and β	Optimal α, β and λ_i
Agreement	67.7%	72.6%▲	75.9%▲

7.6 Conclusions

This chapter focused primarily on proposing evaluation metrics for aggregated search. We introduced a general evaluation framework that captures several traits unique to aggregated search. We instantiated a suite of metrics for evaluating aggregated search pages from this framework. We presented a methodology to collect user preferences over aggregated pages, which allowed us to measure various aspects of our proposed metrics. We did this by simulating aggregated search pages of different quality for a range of topics. The approach allowed us to analyse different parts of the aggregated page pair space. Furthermore, we showed that the proposed metrics correlate well with the majority user preferences and that traditional IR metrics are not well suited to the task. In addition, while some diversity-based metrics can be adapted to measure the preference between page pair, they are not ideal. By instantiating several non-tuned versions of metrics from our framework, we showed that these metrics are at least comparable to diversity-based IR metrics. We also showed that our metrics have the ability to tune their behaviour for pages for which personalised preference data is available.

Although the evaluation metrics we propose have demonstrated to have strong predictive power, we have not compared the effectiveness and weakness of them with existing diversity-aware metrics for evaluating aggregated search. In addition, as we have already known that there are a set of key components in aggregated search, namely, vertical selection, vertical diversity, item selection and result presentation. For all different metrics, we do not understand how discriminative those metrics are and how well different metrics capture and combine different aggregated search key components. Therefore, we further follow-up on the work presented in this chapter by further meta-evaluate the reliability (discriminative power) and intuitiveness in the next chapter.

8 On the Reliability and Intuitiveness of Aggregated Search Metrics

In the previous chapter we focused on presenting a general framework for evaluating the ultimate aggregated search pages and we have demonstrated the proposed metrics have relatively strong predictive power to align with the users. In this chapter, we start our investigations by further meta-evaluating a broad range of evaluation metrics for aggregated search tasks, including aggregated search metrics proposed in the previous chapter, simple single-component aggregated search metrics, adapted diversity IR metrics and traditional IR metrics.

Although several aggregated search (AS) metrics have been proposed to evaluate AS result pages, their properties remain poorly understood. The main differences between these metrics are the way they model each of the four AS compounding factors (vertical selection **VS**, item selection **IS**, result presentation **RP** and vertical diversity **VD**) and combine them. However, how well the metrics capture and combine those factors are poorly understood. This chapter addresses our research questions **RQ 15** to **RQ 16**, as specified in Section 1.2.1. We compare a wide range of AS metrics on two test collections and aim to address the following research questions:

> **RQ 15**: How do all different suites of metrics (traditional IR, diversity IR and aggregated search) perform with respect to reliability, i.e. the ability to statistically discriminate aggregated search systems?
>
> **RQ 16**: Are all different suites of metrics perform sufficiently intuitive to capture different key components of aggregated search?

Our main criteria of comparison are reliability and intuitiveness. By *reliability*, we mean the ability of a metric to detect "actual" performance differences as opposed to those observed by chance, and by *intuitiveness*, we mean the ability to capture any property deemed important to a factor. We focus on how the metrics reflect the four AS compounding factors, **VS**, **IS**, **RP** and **VD**. In this chapter, reliability is measured using discriminative power [Sakai, 2006]. We use the randomised Tukey's Honestly Significant Differences test [Carterette, 2012] because, as shown by [Sakai, 2012], this test is less likely to find significant differences that are not "actual". For intuitiveness, we quantitatively measure the preference agreements using concordance [Sakai, 2012] of a given AS metric with a "basic" single-component metric for each of the four AS factors. Finally, to gain further understanding, we examine AS metrics' ability to capture combination of key components. Our study shows that the AS metrics that capture key AS components (e.g., vertical selection) have several advantages over other metrics.

Our work [Zhou et al., 2013b] is the first endeavour to study the reliability and intuitiveness of AS metrics. We also present an examination of an extensive set of metrics, including a comprehensive set of adapted diversity metrics. Furthermore, we employ the use of two AS collections that form the basis of our results. We found that in terms of reliability (discriminative power), our proposed aggregated search metrics proposed in the previous chapter and α-$nDCG$ [Clarke et al., 2008] are the most discriminative metrics. In terms of intuitiveness to capture single aggregated search factor, there is no single metric that is superior to all other metrics. However, our proposed aggregated search metric AS_{RBP} [Zhou et al., 2012c] in the previous chapter is found to be the most intuitive metric to emphasise all AS components. This work sheds new lights on the further developments and applications of aggregated search metrics.

The remainder of this chapter is organized as follows. Section 8.1 discusses previous work. We describe the metrics investigated in this study in Section 8.2, which also contains the description of our "meta-evaluation" methodology. Details of the data sets and experimental set-up are provided in Section 8.3. Section 8.4 reports our experimental results. We conclude the work in Section 8.5.

8.1 Previous Work

Section 8.1.1 reviews both traditional and diversity metrics used in IR. In this chapter, we adapt some of them to AS. Section 8.1.2 provides an overview of existing work on evaluating AS, either by measuring the performance of key components in isolation or as a whole. Section 8.1.3 summarised methodologies that have been used to compare metrics.

8.1.1 IR Metrics

Traditional Metrics

Traditional IR evaluation is based on topical relevance, $qrel(q, d)$, between a query q and a document d. Traditional IR metrics ignore the document type (e.g. vertical) and measure the quality of a ranked list l by modelling the gain $G@l$ of a user reading all documents in that list l. For instance, $P@k$ assumes that after reading the top k results in l, a user's gain $G@k$ solely depends on the number of relevant documents within the top k results.

Although this metric is simple and widely used, it does not take into account the ranking position, and furthermore, assumes that relevance is a binary judgement. To incorporate graded relevance and to take a more fine-grained user model into account, $nDCG@l$ was proposed [Järvelin and Kekäläinen, 2002]. By diminishing the impact of lower ranked relevant documents, $nDCG@l$ measures the performance of l by cumulating the diminished gain for each position r. A function $g(r)$ is defined to measure the gain of reading a document. The more relevant the document is, the higher the gain to the user. Finally, the metric score is normalized by an ideal ranked list l^*, obtained by ranking all relevant documents in descending order of their relevance.

$$nDCG@l = \frac{\sum_{r=1}^{l} g(r)/\log(r+1)}{\sum_{r=1}^{l} g^*(r)/\log(r+1)} \tag{8.1}$$

Other metrics have been proposed (e.g. RBP [Moffat and Zobel, 2008], ERR [Chapelle et al., 2009]); the major difference between $nDCG$ and them is the assumed user model and how $g(r)$ is defined.

Diversity Metrics

To consider rewarding topical diversity in ranked lists, a set of diversity metrics have been proposed recently; these include α-$nDCG$[Clarke et al., 2008], IA-$nDCG$ [Agrawal et al., 2009] and $D\#$-$nDCG$ [Sakai and Song, 2011].

α-$nDCG$ extends $nDCG$ to account for diversity by discounting the gains that accrue according to the intent (subtopic) previously encountered in the ranked list. The novelty-biased gain $NG(r)$ is defined as:

$$NG(r) = \sum_i J_i(r)(1-\alpha)^{C_n^{r-1}} \tag{8.2}$$

where $J_i(r) = 1$ if a document at rank r is relevant to the i^th intent and 0 otherwise; $C_i(r) = \sum_{k=1}^{r} J_i(r)$ is the number of documents observed within the top r results that contained the i^{th} intent. The strength of the novelty-biased discount is controlled by α.

Agrawal et al. [Agrawal et al., 2009] apply a traditional measure to each subtopic independently and then combined each value to give the expected value of the measure across all intents. This assumes that for a query q with several intents i, the probability of each intent $P(i|q)$ is available. For example, $nDCG$ for an given intent i ($nDCG_i$) is computed first, and then the intent-aware IA-$nDCG$ is computed as:

$$IA\text{-}nDCG@l = \sum_i P(i|q) nDCG_i@l \tag{8.3}$$

D-$nDCG$ [Sakai and Song, 2011], by analogy to $g(r)$ within $nDCG$, calculates a global gain $GG(r)$ at rank r given various intents:

$$GG(r) = \sum_i P(i|q) g_i(r) \tag{8.4}$$

$g_i(r)$ is the gain value for a document at rank r for intent i. Intent recall I-rec [Zhai et al., 2003], i.e. number of intents covered by a ranked list, can be boosted with the following measure:

$$D\#\text{-}nDCG@l = \gamma I\text{-}rec + (1-\gamma) D\text{-}nDCG@l \tag{8.5}$$

γ controls the trade-off between relevance and diversity.

These metrics were proposed to evaluate the diversity of ranked lists over subtopics, and have been recently adapted to measure AS performance [Zhou et al., 2012c]. We discuss these adaptations next.

8.1.2 AS Metrics

Current AS metrics measure either each AS component (**VS**, **VD**, **IS**, **RP**) in isolation or as a whole. An AS page P is composed of a set of blocks $\{B_1, B_2, ...B_n\}$, where each block B_i consists of a set of items $\{I_{i1}, I_{i2}, ...I_{im}\}$. An item can be a "general web" page or a vertical result.

Vertical selection (**VS**) has been studied in [Arguello et al., 2009, Li et al., 2008, Zhou et al., 2012a,b], where the aim is to measure the quality of the set of selected verticals, compared with an annotated set obtained by collecting manual labels from assessors [Arguello et al., 2009, Zhou et al., 2012a,b] or derived from user interaction data [Li et al., 2008]. The quality is mostly evaluated with standard measures of precision, recall and f-measure [Arguello et al., 2009, Zhou et al., 2012a] using a binary annotated set. Recently, risk has also been incorporated into risk-aware VS metrics [Zhou et al., 2012b].

Recent attempts to evaluate the utility of the whole AS page [Arguello et al., 2011b, Ponnuswami et al., 2011b, Zhou et al., 2012c] consider the three key components **VS**, **IS**, **RP** together. For example, [Ponnuswami et al., 2011b] evaluate the utility of a page based on a user engagement metric (CTR) when user interaction data is available. Others [Bailey et al., 2010] evaluate the utility of the page by asking annotators to make assessments based on a number of criteria (relevance, diversity). Although those works comprehensively evaluate AS pages, it remains costly to gather assessments for all possible AS pages.

[Arguello et al., 2011b] collected pairwise preferences on vertical block-pairs from users, and then measured the AS page quality by calculating the distance between the page in question and the ideal (reference) page; the shorter the distance, the better the page. The ideal reference page is obtained by using a voting method for aggregating all pairwise block preference data into a single ranking.

[Zhou et al., 2012c] followed the Cranfield paradigm and proposed an evaluation framework for measuring AS page quality using two types of assessments, item topical-relevance and vertical-orientation, gathered independently. Topical-relevance assessment $qrel(q, d)$ specifies the topical relevance between a document and a query, whereas vertical-orientation $orient(v_i, q)$ is the fraction of users that prefer a page to contain items from the vertical v_i rather than "general web" results for a query q. An example is $AS_{DCG}(P)$, a metric defined as the expected gain $G(B_i)$ of reading each block B_i on page P divided by the expected effort $E(B_i)$ spent, normalized by the score $AS_{DCG}(P^*)$ of an ideal page P^*:

$$AS_{DCG}(P) = \frac{(\sum_{i=1}^{|P|} 1/\log(B_i)G(B_i))/(\sum_{j=1}^{|P|} 1/\log(B_i)E(B_j))}{(\sum_{i=1}^{|P^*|} 1/\log(B_i)G(B_i))/(\sum_{j=1}^{|P^*|} 1/\log(B_i)E(B_j))} \tag{8.6}$$

The gain $G_(B_i)$ combines vertical-orientation $orient(V_j|W, q)$ and topical-relevance $qrel(I_{ik}|q)$ that relate to the quality of the block in an independent manner:

$$G(B_i) = g(orient(V_j|W, q), \alpha) \times \sum_{k=1}^{|B_i|} qrel(I_{ik}|q) \tag{8.7}$$

where the function $g()$ is used so that the relative gain of the vertical can be altered using a tuning parameter α. The effort of examining a block $E(B_i)$ is defined as the accumulative effort of reading all the items within it, that is $E(B_i) = \sum_{k=1}^{|B_i|} E(I_{ik})$ where the effort $E(I_{ik})$ is assumed to depend on the media type of the item.

Several existing diversity metrics were adapted to evaluate AS in [Zhou et al., 2012c] by treating subtopics as verticals and subtopic importance as vertical-orientation as follows: (i) replacing subtopic importance with $orient(V|W, q)$; (ii) substituting the user model for ranks to a model that applies to blocks; and finally (iii) normalising according to the ideal AS page.

All AS metrics model and combine components of AS (**VS**, **VD**, **IS**, **RP**) differently. We use a subset of them for in-depth analysis of their properties.

8.1.3 Comparing Metrics

To date, and to our knowledge, no existing studies comparing the reliability and usefulness of metrics in the context of AS have been reported. However, this current study is similar to the work by Sakai et al. [Sakai, 2012, Sakai and Song, 2011] and Clarke et al. [Clarke et al., 2011] that compare diversity metrics. We therefore follow a similar methodology. For example, we also use discriminative power [Carterette, 2012] to evaluate AS metrics. The novelty of our contribution lies in the insight that our study brings to the AS area, rather than the more usual linear ranked-list approach. Furthermore, the comprehensive examination on how AS metrics capture and measure the different AS components is both novel and timely.

Discriminative power is not the only way to evaluate an evaluation metric. Indeed, highly discriminative metrics, while desirable, may not necessarily measure everything that we may want *measured*. Recently, Sakai [Sakai, 2012] proposed the intuitiveness test[1] for this exact purpose. The intuitiveness test compares a metric of interest with a simple golden standard metric that captures the most important properties that the metric should satisfy. In our study, we apply the intuitiveness test within the context of AS and define four golden standard metrics, respectively, for the four AS factors **VS**, **VD**, **IS**, **RP**. This allows us to investigate how AS metrics capture the key desirable properties of AS.

We should add that other approaches, especially those relying on human subjects (for instance to assess a metric's predictive power), are important. We have already presented how aggregated search metrics perform with respect to predictive power, in aligning with the user preference. By employing Mechanical Turk users, [Zhou et al., 2012c] and [Sanderson et al., 2010], respectively, examined the predictive power of AS metrics and IR metrics. For example, if a metric prefers one AS page or ranked list over another, does the user also prefer the same page/list? One finding in both works was that AS metrics and IR metrics agree reasonably well with human preferences. Although informative from a user perspective, compared to our study, these studies do not give us much insight into how reliable metrics are at ranking systems, or how well the metrics capture key AS components.

[1] This was later renamed as the concordance test [Sakai and Song, 2013]

Table 8.1: A Suite of Metrics that are Meta-evaluated in this Chapter, including Aggregated Search Metrics, Simple Single-Component Aggregated Search Metrics, Adapted Diversity IR Metrics and Traditional IR Metrics.

Metric	VS	IS	RP	VD	Category
$nDCG$[Järvelin and Kekäläinen, 2002]		✓	✓		Traditional
$P@10$		✓			
α-$nDCG$ [Clarke et al., 2008]		✓	✓	✓	Adapted Diversity
IA-$nDCG$ [Agrawal et al., 2009]	✓	✓	✓	✓	
$D\#$-$nDCG$ [Sakai and Song, 2011]	✓	✓	✓	✓	
AS_{DCG} [Zhou et al., 2012c]	✓	✓	✓	✓	AS
AS_{RBP} [Zhou et al., 2012c]	✓	✓	✓	✓	
AS_{ERR} [Zhou et al., 2012c]	✓	✓	✓	✓	
$prec_v$[Zhou et al., 2012a]	✓				Single component
$mean$-$prec$ [Zhou et al., 2012c]		✓			
$Spearman$-$corr$			✓		
rec_v [Zhou et al., 2012a]				✓	

8.2 Evaluating Evaluation Metrics

We first summarize the AS metrics tested in this study in Section 8.2.1. Sections 8.2.2 and 8.2.3 describe the two methods comparing the "goodness" of AS metrics, using a discriminative power and an intuitiveness test, respectively.

8.2.1 AS Metrics

As discussed in Section 8.1, various AS metrics have been proposed to evaluate key components of AS systems, either in isolation or as a whole. We select a subset of existing AS metrics, listed in Table 8.1. Some metrics incorporate all four factors (**VS**, **IS**, **RP**, **VD**) (e.g. AS_{DCG}) whereas others relate to a subset (e.g. α-$nDCG$). For metrics concerned with the same subset of factors, the way these factors are incorporated can vary. For example, AS_{DCG} and AS_{RBP} mainly vary on their assumed user browsing model so that they give different diminishing returns for documents at latter rank of the page. We also include simple metrics that capture one AS factor (detailed in Section 8.2.3). The selected metrics allow us to investigate all four factors, both individually and when combined, as well the various categories (traditional IR metrics, adapted diversity metrics, AS metrics and simple single-component metrics) to which they belong. Some metrics possess parameters that can be tuned to (de)emphasize a factor (e.g. α in α-$nDCG$ rewards **VD** differently). In this work, we leave properly setting metrics' parameters as future work and will utilise standard parameter settings for each metric (i.e. we follow settings from previous work [Zhou et al., 2012c]).

We briefly explain the differences between the selected metrics (Section 2 has full details). In short, both $nDCG$ and $P@10$ ignore the vertical type and only considers

IS (and **RP** for $nDCG$). Without considering the intent likelihood $P(i|q)$, α-$nDCG$ rewards **VD** by diminishing redundant relevant documents. Although incorporating $P(i|q)$, IA-$nDCG$ considers each intent independently, and was shown to be biased to rewarding relevant documents with high intent (i.e caring less about **VD**) [Clarke et al., 2011]. Comparatively speaking, for each rank, D-$nDCG$ and $D\#$-$nDCG$ accumulate the global gain for all intents, and have been proven to reward **VD** more. The differences between them is that $D\#$-$nDCG$ explicitly boosts **VD** by linearly combining D-$nDCG$ with I-rec. AS_{DCG}, AS_{RBP}, AS_{ERR} reward all four components of AS, but differ in the assumed user browsing model (which affects **RP**).

8.2.2 Discriminative Power

Given a test collection and a set of runs, the discriminative power of a metric is measured by conducting a statistical significance test for every pair of runs, and then counting the number of significant differences. In this chapter, we use the randomised version of Tukey's Honestly Significant Differences (HSD) test [Carterette, 2012]. This test takes the entire set of runs into account when judging the significance of each run pair. This test is more conservative (compared to e.g. bootstrap test [Sakai, 2006]), and hence less likely to lead to significant differences that are not "real". We choose this test because of its reliance on modern computational power instead of statistical assumptions.

Let $t_{T,i}$ denote the i^{th} topic from a topic set T of size N, and $M(t, r_j)$ denotes the value of metric M for topic t and run r_j. The main idea behind Tukey's HSD is that if the largest mean difference observed is not significant, then none of the other differences should be significant either. Given a set of runs, the null hypothesis H_0 is "there is no difference between any of the systems". We perform randomised Tukey's HSD as shown in Algorithm 1 (taken from [Carterette, 2012]). From a given matrix X whose element at (row i, column j) represents the performance of the j^{th} run for the i^{th} topic, we create B new matrices X^b by permuting each row at random; then, for every run pair, we compare the performance δ of this run pair with the largest performance δ observed within X^b. Finally, for each run pair, we obtain the Achieved Significance Level (ASL or p-value), which represents how likely this would be under H_0 (null hypothesis), is computed for each run pair (see Algorithm 1). As in any other significance test, H_0 is rejected if $ASL < \alpha$.

Using the results of the randomised Tukey's HSD tests, we also try to estimate the performance δ required to achieve a statistical significance at α for a given topic set size as shown in Algorithm 2: we simply take the smallest observed δ from all the run pairs that were found to be significantly different.

8.2.3 Intuitiveness

We now discuss concordance test that examine the intuitiveness of metrics. AS metrics aim to balance the four key AS factors (**VS**, **IS**, **RP** and **VD**) when assessing performance. Inevitably, they tend to be complex, making it particularly difficult to determine if a metric is "measuring what we want to measure". To address this, Sakai [Sakai, 2012] proposed a simple method for quantifying "which metric is more intuitive", and this

```
foreach pair of runs ($r_1$, $r_2$) do count($r_1$, $r_2$) = 0; ;
for $b = 1$ to $B$ do
    create matrix $X^{*b}$ whose row $t$ is a permutation of row $t$ of $X$ for every $t \in T$;
    $max^{*b} = max_i \bar{x}_i^{*b}$; $min^{*b} = min_i \bar{x}_i^{*b}$ where
    $x_i^{*b}$ is the mean of i-th column vector of $X^{*b}$;
    foreach pair of runs ($r_1$, $r_2$) do
        if $max^{*b} - min^{*b} > |\bar{x}(r_1) - \bar{x}(r_2)|$ then where
        $\bar{x}(r_i)$ is the mean of column vector for $r_i$ in $X$
            $count(r_1, r_2) + +$;
foreach pair of runs ($r_1$, $r_2$) do
    $ASL(r_1, r_2) = count(r_1, r_2)/B$;
```

Algorithm 1: Obtaining the Achieved Significance Level with the two-sided, randomised Tukey's HSD given a performance value matrix X whose rows represent topics and columns represent runs.

```
foreach pair of runs $(r_1, r_2)$ with a significant difference at $\alpha$ do
    $\delta_\alpha(r_1, r_2) = |mean(r_1) - mean(r_2)|$ ;
$\delta_\alpha = min_{i,j} \delta_\alpha(r_1, r_2)$ ;
```

Algorithm 2: Estimating the performance δ required for obtaining a significant difference at α with the randomised Tukey's HSD test.

has been applied to measuring intuitiveness for diversity IR metrics. We now apply his approach to AS.

The concordance test algorithm [Sakai, 2012] is shown in Algorithm 3. The algorithm computes relative concordance scores for a pair of metrics M_1 and M_2 and a gold standard metric M_{GS}. The latter represents a basic property that a candidate metric should satisfy. For our study, we consider four simple metrics as our gold standards, one for each AS factor. Note that these gold standards are simple and some of them (e.g. **VS**, **VD**, **IS**) are set retrieval metrics based on binary relevance. Since different AS metrics employ different position-based discounting and different ways to define graded topical relevance, the gold standards should be as agnostic to these differences as possible. Their purpose is to separate out and test the important properties of the more complex AS metrics. The four gold standard metrics are:

- Simple **VS** metric: vertical precision $prec_v$.
- Simple **VD** metric: vertical recall rec_v.
- Simple **IS** metric: mean precision $mean\text{-}prec$ of vertical result items.
- Simple **RP** metric: Spearman's rank correlation $corr$ with "perfect" AS reference page.

For a vertical V_i and query q, we consider the vertical to be relevant if $orient(V_i|W, q)$ is greater than 0.5[2]. Note that the vertical recall rec_v can also be refered as $I\text{-}rec$ (intent

[2] We select the threshold as 0.5 as 50% assessors majority preference is a suitable percentage whereby the assessments are neither too noisy (25%) or stringent (100%) as we are also interested in the vertical diversity.

```
$Disagreements = 0; Correct_1 = 0; Correct_2 = 0;$
foreach pair of runs $(r_1, r_2)$ do
    foreach topic $t$ do
        $\delta M_1 = M_1(t, r_1) - M_1(t, r_2);$
        $\delta M_2 = M_2(t, r_1) - M_2(t, r_2);$
        $\delta M_{GS} = M_{GS}(t, r_1) - M_{GS}(t, r_2);$
        if $(\delta M_1 \times \delta M_2) < 0$ then // $M_1$ and $M_2$ disagree
            $Disagreements + +;$
            v if $\delta M_1 \times \delta M_{GS} \geq 0$ then // $M_1$ and $M_{GS}$ agree
                $Correct_1 + +;$
            if $\delta M_2 \times \delta M_{GS} \geq 0$ then // $M_2$ and $M_{GS}$ agree
                $Correct_2 + +;$
$Intuitive(M_1|M_2, M^{GS}) = Correct_1/Disagreements;$
$Intuitive(M_2|M_1, M^{GS}) = Correct_2/Disagreements;$
```

Algorithm 3: Computing the concordance of metrics M_1 and M_2 based on preference agreement with M_{GS}.

recall). Moreover, simple RP metric *corr* is similar to voting approach from [Arguello et al., 2011b]. However, rather than rewarding more on higher positions of the page, *corr* calculates the correlation by weighting each position equally.

The steps conducting the concordance test are as follows: We first obtain all pairs of AS systems/pages for which M_1 and M_2 disagree with each other. Then, out of these disagreements, we count how often each metric agrees with the gold standard metric. In this way, we can discuss which of the two metrics is the most "intuitive". Moreover, we can argue that an ideal metric should be consistent with all four gold standards; we therefore add one additional step by counting how often the metric agrees with an essential subset of or all four gold standards.

8.3 Data

To provide findings not tailored to one data set, and hence generalisable, our experiments are conducted on the two test collections described in Section 8.3.1. The methodology employed to simulate AS systems is presented in Section 8.3.2.

8.3.1 Test collections

An AS test collection consists of a number of verticals, each populated by items of that vertical type, a set of topics expressing information needs relating to one or more verticals, and assessments indicating both the topical-relevance of the items and the perceived user-oriented usefulness of their associated verticals to each of the topics.

The first test collection is an AS test collection [Zhou et al., 2011] created by reusing an existing web collection (as described in Chapter 4), ClueWeb09. The verticals were

In addition, the relevant vertical set obtained from this simple thresholding approach is similar to that obtained from [Arguello et al., 2011b] voting approach (where more relevance assessments are needed).

created either by classifying items in the web collections into different genres (e.g. blog, news) or by adding items from existing multimedia collections (e.g. image, video). The topics and topical-relevance assessments of the items across the verticals were obtained by reusing the assessments developed in two TREC evaluation tasks (TREC Web Track and Million-Query Track). The verticals used are listed in Table 4.1 in Chapter 4, and correspond to real-world usage of verticals by commercial search engines.

The second AS test collection [Nguyen et al., 2012] is a new dataset that will be used in TREC FedWeb track 2013[3]. The collection contains search result pages from 108 web search engines (such as Google, Yahoo!, YouTube and Wikipedia). For each engine, several query-based samplings were provided for vertical selection. Relevance judgements were collected by judging both the snippet created by the engine, and the actual document content for the results returned by the engines for a set of queries (reused TREC Web Track 2010 queries). To use the same verticals as listed in Table 3.2, we manually mapped the 108 search engines into them. This was straightforward since the engine categories used were similar to those in Table 4.1.

The vertical-orientation information of each topic from the first test collection was obtained by providing the vertical names (with a description of their characteristics) and asking a set of assessors to make pairwise preference assessments, comparing each vertical in turn to the reference "general web" vertical ("is adding results from this vertical likely to improve the quality of the ten blue links?") [Zhou et al., 2012a]. Note that since the two test collections contain the same set of topics (reused from TREC Web Track 2010), the vertical-orientation information from the first collection can be used for the second collection. Some details and statistics of the two test collections are shown in Table 8.2.

8.3.2 Simulating AS System Runs

For each topic, we simulate a set of aggregated search pages. We assume that a page consists of ten "general web" blocks (one "general web" page is a block) and up to three vertical blocks dispersed throughout those ten blocks (where each vertical block consists of a fixed number of three items). Recall that there are three key components of an aggregated search system that can be varied: (i) Vertical Selection (**VS**) (ii) Item Selection (**IS**) and (iii) Result Presentation (**RP**). We generate pages by simulating an aggregated search system in which the three components vary in quality.

We simulate four different state-of-the-art VS strategies, namely *ReDDE* [Si and Callan, 2003a], *CRCS(e)* [Shokouhi, 2007], *click-through* [Arguello et al., 2011a] and *vertical-intent* [Arguello et al., 2011a]. Deriving from sampled vertical representation, *ReDDE* and *CRCS(e)* model each verticals average document score in a full-dataset retrieval (all sources together). By contrast, *click-through* and *vertical-intent* use, respectively, users' click-through data and issued queries from a search engine log (AOL-log). Similar to [Arguello et al., 2011a], we model VS as a classification task and, for each single VS approach (e.g. ReDDE), the output is n independent prediction probability scores (one per vertical, n is the number of verticals).

Assuming four vertical positions (ToP, MoP, BoP, None) on the page, each candidate

[3] https://sites.google.com/site/trecfedweb/

Table 8.2: Two Test Collections Used to Meta-evaluating Aggregated Search Metrics with respect to Reliability and Intuitiveness.

Test Collection (a)	classified TREC 2009 Web Track ClueWeb Category B ("VertWeb11") [Zhou et al., 2011]
Documents	ClueWeb09 Cat B (approximately 50 million documents)
Topics	a subset of 56 topics (pertaining various vertical intents) is used in our experiments, selected from 320 topics (reused TREC 2009-2010 Web Track and TREC 2008-2009 MillionQuery Track topics)
Intents	12 vertical intents (with an average of 1.83 relevant verticals per topic)
Runs	36 simulated AS systems

Test Collection (b)	TREC 2013 FedWeb Track Data ("FedWeb13") [Nguyen et al., 2012]
Documents	Documents sampled from 108 heterogeneous search engines
Topics	50 topics (TREC 2010 Web Track) that cover multi-interpretations or multiple facets
Intents	12 vertical intents manually classified for 108 search engines
Runs	36 simulated AS systems

vertical prediction is compared with three threshold parameters γ_{1-3} (one for each position) to assign the corresponding embedding position. A given vertical is assigned to the highest position for which the vertical prediction probability is greater than or equal to all thresholds below it and verticals within the same position are ordered by descending order of prediction probability. Using similar techniques in [Arguello et al., 2011b], we obtained a separate development set to tune the three threshold parameters.

For **IS** we simulate three potentially different levels of relevance. These are *Perfect*, *BM25*, and *TF*. *Perfect* selects all items in the vertical that are topically relevant. *BM25* and *TF* select the top three ranked items from the rankings provided by the BM25 and a simple TF (term-frequency) weighting respectively, with the PageRank score as a prior for both *BM25* and *TF*.

For **RP**, we simulate three different result presentation approaches: *Perfect*, *Random* and *Bad*. *Perfect* places the vertical blocks on the page so that gain could potentially be maximised, i.e. all the relevant items are placed before non-relevant items. However, if these items are part of a vertical, we position the highest orientated vertical first. *Random* randomly disperses the vertical blocks on the page while maintaining the position of the "general web" blocks. *Bad* reverses the perfectly presented page.

By varying the quality of each of the three key components, we can vary the quality of the result pages created by an aggregated search system in a more controlled way. For each topic, we can create 36 (4 × 3 × 3) system runs[4]. Therefore, for the discriminative power test, we have C_{36}^{2} (630) system pairs. Using this approach we can create a near ideal aggregated page for a query by using *Perfect* **VS**, *Perfect* **IS**, and *Perfect* **RP**. This

[4] Certain combinations of VS, IS, and RP do not create unique simulated pages.

is a greedy approach to the problem and is used as our method of normalisation.

8.4 Experiments

We experiment with both discriminative power (Section 8.4.1) and intuitiveness (Section 8.4.2) of AS metrics.

8.4.1 Disciminative Power Results

Using the two AS test collections (VertWeb11 and FedWeb13), we evaluated two set of metrics in terms of discriminative power. The first set consists of metrics that evaluate only a subset of components within AS: $nDCG$, $P@10$ (Traditional Metrics), $prec_v$, rec_v, $mean\text{-}prec$ and $corr$ (Single-Component Metrics as described in Section 3.3). The second set presents the most extensive set of AS metrics and adapated diversity metrics), including (recently proposed AS metrics [Zhou et al., 2012c] AS_{DCG}, AS_{RBP}, AS_{ERR} and adapted diversity metrics $\alpha\text{-}nDCG$, $IA\text{-}nDCG$ and $D\#\text{-}nDCG$. Note that we used the standard parameter settings in this experiments (e.g. setting $\alpha = 0.5$ for $\alpha\text{-}nDCG$, etc.) and extensive tunning of metrics' parameters are left for future work.

Figures 8.1 and 2 show the ASL curves of some selected AS metrics, based the randomised Tukey's HSD on FedWeb13 and VertWeb11 collection respectively. Part (a) of each figure (higher part) shows the results with the first set of metrics (traditional IR and AS component-based metrics), to discuss which subset of components can be more discriminative on ranking AS systems. Part (b) of each figure (lower part) shows an extensive set of AS metrics used in the literature (as described in Section 2). With different approaches in modelling and combining components, the idea is to get insights on how they perform on discriminating AS systems. The metrics that are more discriminative are those closer to the origin in the figures. Table 8.3 cut those two figures in half at $\alpha = 0.05$ to quantify discriminative power and the performance δ required for achieving statistical significance with a given number of topics (56 for VertWeb11 or 50 topics for FedWeb13). For example, left side of Table 8.3(a) shows that the discriminative power of component-based metrics $mean\text{-}prec$ according to the Tukey's HSD test at $\alpha = 0.05$ is (125/630) = 19.8% (i.e., 100 significantly different run pairs were found) and the δ required for achieving statistical significance is around 0.12.

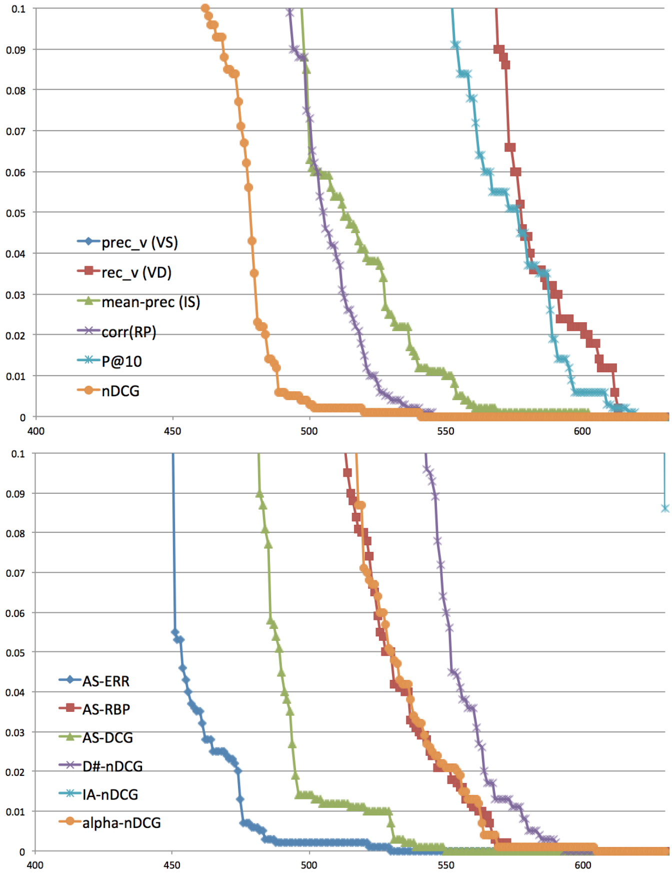


Figure 8.1: FedWeb13 Discriminative Power Evaluation: ASL curves based on the randomised Tukey's HSD. y-axis: ASL (i.e., p-value); x-axis: run pairs sorted by ASL.

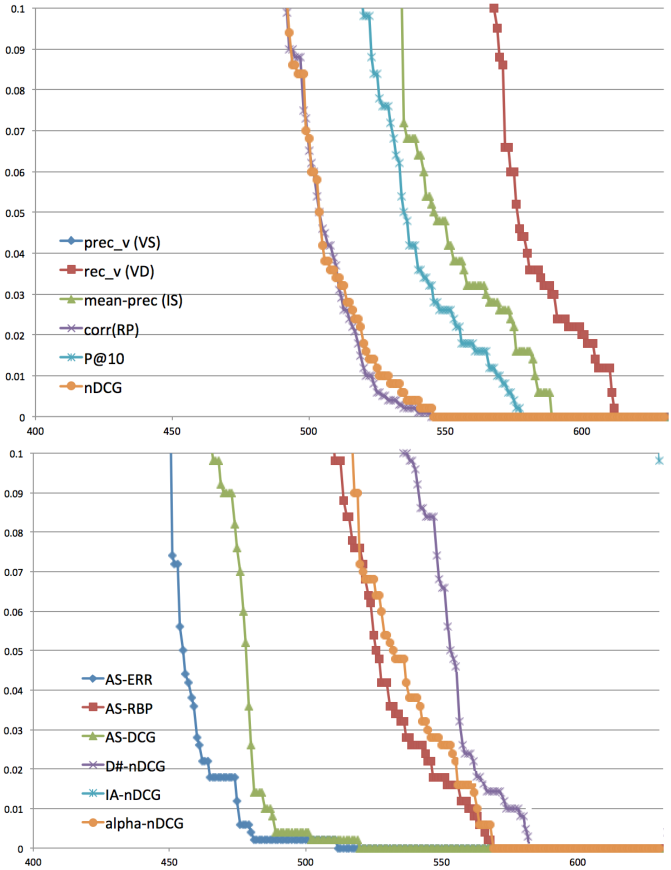


Figure 8.2: VertWeb11 Discriminative Power Evaluation: ASL curves based on the randomised Tukey's HSD. y-axis: ASL (i.e., p-value); x-axis: run pairs sorted by ASL.

Table 8.3: Discriminative power / performance δ of metrics (single-component and AS metrics) based on the randomised Tukey's HSD test at $\alpha = 0.05$ on FedWeb13 and VertWeb11 collections.

(a) FedWeb13						(b) VertWeb11					
$prec_v$	0.0%	N/A	α-$nDCG$	15.9%	0.08	$prec_v$	0.0%	N/A	α-$nDCG$	14.5%	0.08
rec_v	8.3%	0.09	IA-$nDCG$	0.0%	N/A	rec_v	7.9%	0.10	IA-$nDCG$	0.0%	N/A
$mean$-$prec$	19.8%	0.12	$D\#$-$nDCG$	12.5%	0.07	$mean$-$prec$	13.0%	0.12	$D\#$-$nDCG$	12.1%	0.07
$corr$	20.8%	0.11	AS_{DCG}	22.4%	0.10	$corr$	20.4%	0.10	AS_{DCG}	23.8%	0.12
$nDCG$	24.1%	0.10	AS_{RBP}	15.9%	0.05	$nDCG$	20.5%	0.13	AS_{RBP}	15.3%	0.05
P@10	8.6%	0.13	AS_{ERR}	28.1%	0.09	P@10	14.9%	0.10	AS_{ERR}	27.7%	0.09

Let "$M_1 \in M_2$" denote the relationship: "M_2 outperforms M_1 in terms of discriminative power." First, by comparing the different component-based metrics in terms of discriminative power as shown in Part (a) (higher) Figures 1 and 2, and left side of Table 8.3(a) and (b) , the following trends can be observed: $prec_v \in rec_v \in$ ($mean$-$prec$, P@10) $\in$ ($nDCG$, $corr$). We summarise the interesting findings as below:

- Single-component metrics perform comparatively well on discriminating AS systems. RP metric $corr$ appears to be the most consistently discriminative metrics of all the single-component metrics for our data sets, achieving discriminative power comparable to traditional IR metrics (e.g. P@10 or $nDCG$).

- VS metric $prec_v$ is the least discriminative single component metric for evaluating AS pages. After a close examination, we found that since most of AS pages only present a few verticals (mostly 1 or 2), the possible values of $prec_v$ are quite limited across pages and therefore it is not discriminative.

- For traditional IR metric, $nDCG$ performs consistently better than P@10 and other single-component metrics. It is not surprising $nDCG$ performs better than P@10 since it incorporates both ranking position and graded relevance assessments. However, it is interesting to observe that, without considering rank-based discount, $corr$ is able to discrminate AS systems comparably to $nDCG$ in the VertWeb11 collection.

Next, as shown in Part (b) (lower) in Figures 1 and 2, and right side of Table 8.3(a) and (b), we compare the different AS metrics in terms of discriminative power and we observe: IA-$nDCG \in D\#$-$nDCG \in$ (AS_{RBP}, α-$nDCG$) $\in AS_{DCG} \in AS_{ERR}$. The interesting findings are summarised as below:

- AS-metrics (e.g. AS_{ERR}) are generally more discriminative than other adapted diversity metrics. This is not surprising since AS-metrics incorporates effort of reading different multi-media items and they might be more powerful/discriminative in controling reward trade-off between vertical orientation and topical relevance.

- AS_{ERR} outperforms over AS_{DCG} and AS_{RBP} on discriminative power. When considering graded relevance assessments, this might suggest that metrics considering inter-dependency of relevance among documents can perform more discriminating than other approaches that only discounted over positions.

- IA-$nDCG$ and $D\#$-$nDCG$ performs least discriminative over two data sets. IA-$nDCG$'s failure might be explained by its top-heavy characteristic (i.e. it heavily rewards highly oriented vertical results). $D\#$-$nDCG$'s discrimination might be affected by the less discriminating rec_v that it has incorporated.

- Generally, different AS metrics are more discriminative than single-component metrics and traditional IR metrics (although with the exception of $corr$ and $nDCG$ being more discriminative than e.g. $D\#$-$nDCG$).

8.4.2 Intuitiveness Results

Highly discriminative metrics, while desirable, may not necessarily measure what we want to measure. The aim of this section is to answer the question: how do the different AS metrics differ from one another, and which ones are more intuitive than others for the purpose of search result aggregation? We answer this by conducting concordance test.

Table 8.4 and 8.5 show the "intuitiveness" scores for a variety of AS metrics, computed using the preference agreement algorithm shown in Algorithm 3. As specified in Table 8.1, our tested AS metrics include a variety of adapted diversity metrics, existing AS metrics and a set of single-component AS metrics. Specifically, we select an essential set of metrics (i.e. α-nDCG, IA-$nDCG$, AS_{DCG}, $D\#$-$nDCG$) that represent different frameworks in modelling AS evaluation. In addition, in order to provide insights on the effectiveness of different user models (e.g. position-based discount model, cascade model) that are utilised for AS evaluation , we also include AS_{DCG}, AS_{RBP} and AS_{ERR} and investigate their ability to capture different key AS components. Here, due to space limitation, we only show results in Table 8.4 and 8.5 for FedWeb13 collection since we found the results are similar across our two test collections. Note that as we have $36 * 35/2 = 630$ run pairs, we have $50 * 630 = 31500$ pairs of aggregated search pages for the tests.

As specified above on testing intuitiveness on four AS components (**VS**, **IS**, **RP** and **VD**), part (a) uses precision of returned vertical set $prec_v$ as the gold-standard and therefore represents how the AS metrics favour aggregated pages that select majority-preferred (i.e. relevant) verticals; Part (b) uses recall of verticals rec_v as the gold-standard and therefore represents how they favour the result with more diverse sets of verticals; Part (c) computes the intuitiveness scores by showing how AS metrics favors the returned set that contain large number of relevant documents, as measured by the mean of precision for each vertical results; and part (d) measures the "goodness" of AS systems embeding vertical results into "general web" results and utilizes the Spearman Rank Correlation between the AS page of interest and the reference AS page ("perfect" page) as the measure. For example, Table 8.4 (a) shows that, for FedWeb13 collection, if we compare α-$nDCG$ and IA-$nDCG$ in terms of component of **VS** (i.e. the ability to select relevant verticals), there are 10222 disagreements, and that while the intuitive score for α-$nDCG$ is only 0.742, that for IA-$nDCG$ is 0.792. This means that, given a pair of AS pages for which α-$nDCG$ and IA-$nDCG$ disagree with each other, IA-$nDCG$ is more likely to agree with $prec_v$ on the preference than α-$nDCG$.

Let "$M_1 > M_2$" denote the relationship: "M_1 statistically significantly outperforms M_2 in terms of concordance with a given gold-standard metric." As we assume the used simple single-component metrics can properly reflect the performance of each component, when comparing different frameworks (i.e. α-$nDCG$, IA-$nDCG$, $D\#$-$nDCG$ and AS_{DCG}) for capturing individual key AS component, several trends can be observed relatively from Table 8.4 as follows [5]:

- Concordance with $prec_v$ (pure vertical orientation): IA-$nDCG$ $> AS_{RBP} > AS_{DCG}$ $> D\#$-$nDCG$ $> AS_{ERR}$, α-$nDCG$;

[5] In general, note that pairwise statistical significance is not transitive. However, it turns out that our results do not violate transitivity.

Table 8.4: Concordance Test Results with the TREC FedWeb13 Track Data on Capturing Single Aggregated Search Component (50 topics; 36 simulated runs). Statistically Significant Differences with the Sign Test are indicated by △($\alpha = 0.05$) and ▲($\alpha = 0.01$).

(a). **(VS)** gold standard: vertical selection precision ($prec_v$)					
	IA-$nDCG$	$D\#$-$nDCG$	AS_{DCG}	AS_{RBP}	AS_{ERR}
α-$nDCG$	0.742/**0.792**△ (10222)	0.755/**0.783**△ (9882)	0.763/**0.784**△ (8857)	0.758/**0.788**△ (9155)	0.769/**0.783** (10586)
IA-$nDCG$	-	**0.798**/0.713▲ (6105)	**0.797**/0.696▲ (6521)	**0.775**/0.715▲ (7595)	**0.802**/0.715▲ (8764)
$D\#$-$nDCG$	-	-	0.747/**0.770**△ (5664)	0.751/**0.759** (6262)	**0.781**/0.754△ (8453)
AS_{DCG}	-	-	-	0.728/**0.784**△ (3230)	**0.771**/0.754 (5351)
AS_{RBP}	-	-	-	-	**0.775**/0.742△ (8309)
(b). **(VD)** gold standard: vertical recall (rec_v)					
	IA-$nDCG$	$D\#$-$nDCG$	AS_{DCG}	AS_{RBP}	AS_{ERR}
α-$nDCG$	0.616/**0.875**▲ (10222)	0.605/**0.881**▲ (9882)	0.660/**0.831**▲ (8857)	0.664/**0.831**▲ (9155)	0.653/**0.828**▲ (10586)
IA-$nDCG$	-	0.672/**0.747**▲ (6105)	**0.871**/0.531▲ (6521)	**0.863**/0.571▲ (7595)	**0.847**/0.605▲ (8764)
$D\#$-$nDCG$	-	-	**0.917**/0.442▲ (5664)	**0.920**/0.493▲ (6262)	**0.874**/0.568▲ (8453)
AS_{DCG}	-	-	-	0.726/**0.735** (3230)	0.742/**0.761** (5351)
AS_{RBP}	-	-	-	-	0.732/**0.741** (8309)
(c). **(IS)** gold standard: mean precision of vertical retrieved items ($mean$-$prec$)					
	IA-$nDCG$	$D\#$-$nDCG$	AS_{DCG}	AS_{RBP}	AS_{ERR}
α-$nDCG$	0.358/**0.838**▲ (10222)	0.314/**0.883**▲ (9882)	0.331/**0.867**▲ (8857)	0.312/**0.892**▲ (9155)	0.408/**0.791**▲ (10586)
IA-$nDCG$	-	0.430/**0.779**▲ (6105)	0.526/**0.665**▲ (6521)	0.466/**0.750**▲ (7595)	**0.738**/0.454▲ (8764)
$D\#$-$nDCG$	-	-	**0.686**/0.470▲ (5664)	0.598/**0.603** (6262)	**0.843**/0.292▲ (8453)
AS_{DCG}	-	-	-	0.412/**0.801**▲ (3230)	**0.897**/0.255▲ (5351)
AS_{RBP}	-	-	-	-	**0.857**/0.293▲ (8309)
(d). **(RP)** gold standard: Spearman Correlation with “perfect” AS page ($corr$)					
	IA-$nDCG$	$D\#$-$nDCG$	AS_{DCG}	AS_{RBP}	AS_{ERR}
α-$nDCG$	**0.640**/0.504▲ (10222)	**0.625**/0.527▲ (9882)	**0.618**/0.537▲ (8857)	**0.620**/0.535▲ (9155)	**0.601**/0.544△ (10586)
IA-$nDCG$	-	0.469/**0.640**▲ (6105)	0.438/**0.657**▲ (6521)	0.466/**0.640**▲ (7595)	0.463/**0.673**▲ (8764)
$D\#$-$nDCG$	-	-	0.511/**0.587**▲ (5664)	0.538/**0.585**△ (6262)	0.515/**0.617**▲ (8453)
AS_{DCG}	-	-	-	**0.585**/0.543△ (3230)	0.553/**0.632**▲ (5351)
AS_{RBP}	-	-	-	-	0.544/**0.611**▲ (8309)

- Concordance with rec_v (pure vertical diversity): $D\#\text{-}nDCG$ >$IA\text{-}nDCG$ >AS_{DCG}, AS_{RBP}, AS_{ERR} >$\alpha\text{-}nDCG$;
- Concordance with $mean\text{-}prec$ (pure item topical relevance): AS_{RBP}, $D\#\text{-}nDCG$ >AS_{DCG} >$IA\text{-}nDCG$ >AS_{ERR} >$\alpha\text{-}nDCG$;
- Concordance with $corr$ (presentation): $\alpha\text{-}nDCG$ >AS_{ERR} >AS_{DCG} >AS_{RBP} >$D\#\text{-}nDCG$ >$IA\text{-}nDCG$.

To summarise, (1). The intent-aware (IA) [Agrawal et al., 2009] and recently proposed AS-metric evaluation framework [Zhou et al., 2012c] works best for rewarding selecting relevant verticals based on the intuitiveness score; (2). $D\#$ and IA framework favors rewarding on vertical diversity (promoting diverse set of results from different verticals). It is not surprisingly that $D\#$ framework behave similar to rec_v since $D\#$ boosts diversity by incorporating $I\text{-}rec$ into the framework. (3). $D\#$ and AS metrics tends to reward more result page with more topically relevant items while AS_{RBP} works best, compared to other user models; (4). $\alpha\text{-}nDCG$ and AS_{ERR} consistently perform worst on vertical orientation (VS), vertical diversity (VD) and topical relevance (IS). (5). $\alpha\text{-}nDCG$ and AS_{ERR} are better correlated with result presentation (RP) evaluation. From a close examination, this is because the cascade model can better discriminate the small relevance “exchanges” (differences) on top or bottom of the page.

To investigate how the above metrics accurately combine components, we also conduct concordance test, to answer: how often does a given metric agree with a set of components at the same time? We are reporting in Table 8.5 the interesting component combinations that represent the most crucial aspects of AS, i.e. VS+IS (orientation and relevance), VS+IS+VD (orientation, relevance and diversity) and VS+IS+RP+VD (ultimate utility). The results can be summarised as below:

- Concordance with $prec_v$ AND $mean\text{-}prec$ (vertical orientation and topical relevance): AS_{RBP} >$D\#\text{-}nDCG$ >AS_{DCG} >$IA\text{-}nDCG$ >AS_{ERR} >$\alpha\text{-}nDCG$;
- Concordance with $prec_v$ AND rec_v AND $mean\text{-}prec$ (vertical orientation, topical relevance and diversity): $D\#\text{-}nDCG$ >AS_{RBP}, $IA\text{-}nDCG$ >AS_{DCG} >AS_{ERR} >$\alpha\text{-}nDCG$;
- Concordance with all: AS_{RBP} >$D\#\text{-}nDCG$ >AS_{DCG}, $IA\text{-}nDCG$ >AS_{ERR} >$\alpha\text{-}nDCG$.

To summarise, we find that $D\#\text{-}nDCG$, AS_{RBP} performs best on combining components while $D\#$ metric captures better on (vertical diversity (VD) and AS_{RBP} models better on vertical orientation and relevance (VS, IS). Moreover, we quantitatively show the advantages of metrics that capture key components of AS (e.g. VS) over those that do not (e.g. $\alpha\text{-}nDCG$).

8.5 Conclusions

This chapter focused primarily on proposing to measure performance of AS metrics based on both discriminative power and intuitiveness. To our knowledge, this is the

Table 8.5: Concordance Test Results with the TREC FedWeb13 Track Data on Capturing Multiple Aggregated Search Components (50 topics; 36 simulated runs). Statistically Significant Differences with the Sign Test are indicated by $\triangle(\alpha = 0.05)$ and $\blacktriangle(\alpha = 0.01)$.

(VS+IS) gold standard: vertical selection precision AND vertical item mean precision ($prec_v$+$mean$-$prec$)

	IA-$nDCG$	$D\#$-$nDCG$	AS_{DCG}	AS_{RBP}	AS_{ERR}
α-$nDCG$	0.253/**0.656▲** (10222)	0.237/**0.689▲** (9882)	0.256/**0.677▲** (8857)	0.236/**0.701▲** (9155)	0.317/**0.616▲** (10586)
IA-$nDCG$	-	0.351/**0.541▲** (6105)	0.424/**0.443** (6521)	0.352/**0.524▲** (7595)	**0.598**/0.301▲ (8764)
$D\#$-$nDCG$	-	-	**0.531**/0.348▲ (5664)	0.431/**0.454△** (6262)	**0.661**/0.213▲ (8453)
AS_{DCG}	-	-	-	0.267/**0.632▲** (3230)	**0.700**/0.188▲ (5351)
AS_{RBP}	-	-	-	-	**0.672**/0.200▲ (8309)

(VS+IS+VD) gold standard: vertical selection precision AND vertical item mean precision AND vertical recall ($prec_v$+$mean$-$prec$+rec_v)

	IA-$nDCG$	$D\#$-$nDCG$	AS_{DCG}	AS_{RBP}	AS_{ERR}
α-$nDCG$	0.195/**0.593▲** (10222)	0.180/**0.619▲** (9882)	0.218/**0.594▲** (8857)	0.203/**0.612▲** (9155)	0.267/**0.545▲** (10586)
IA-$nDCG$	-	0.257/**0.421▲** (6105)	**0.392**/0.271▲ (6521)	0.327/**0.347** (7595)	**0.541**/0.205▲ (8764)
$D\#$-$nDCG$	-	-	**0.495**/0.179▲ (5664)	**0.410**/0.273▲ (6262)	**0.596**/0.128▲ (8453)
AS_{DCG}	-	-	-	0.237/**0.527▲** (3230)	**0.576**/0.172▲ (5351)
AS_{RBP}	-	-	-	-	**0.552**/0.179▲ (8309)

(VS+IS+RP+VD) golden standard: ALL single-component metrics ($prec_v$+$mean$-$prec$+rec_v+$corr$)

	IA-$nDCG$	$D\#$-$nDCG$	AS_{DCG}	AS_{RBP}	AS_{ERR}
α-$nDCG$	0.131/**0.350▲** (10222)	0.117/**0.368▲** (9882)	0.139/**0.361▲** (8857)	0.128/**0.369▲** (9155)	0.164/**0.332▲** (10586)
IA-$nDCG$	-	0.128/**0.263▲** (6105)	**0.194**/0.183 (6521)	0.166/**0.235▲** (7595)	**0.287**/0.137▲ (8764)
$D\#$-$nDCG$	-	-	**0.248**/0.090▲ (5664)	0.162/**0.211△** (6262)	**0.324**/0.071▲ (8453)
AS_{DCG}	-	-	-	0.131/**0.315▲** (3230)	**0.345**/0.113▲ (5351)
AS_{RBP}	-	-	-	-	**0.326**/0.104▲ (8309)

first study to extensively examine properties of metrics in the context of AS. We used most extensive set of existing AS metrics and adapted diversity metrics and test them across two AS test collections.

Our main experimental findings are:

- In terms of discriminative power, (a). In terms of four components of AS, RP ($corr$) is the most discriminative feature (metric) for evaluation, followed by IS ($mean$-$prec$), VD (rec_v) and VS ($prec_v$); (b). AS-metrics (e.g. AS_{ERR}, AS_{DCG}) and α-$nDCG$ are the most discriminative metrics and are superior to $D\#$-$nDCG$ and IA-$nDCG$.

- In terms of intuitiveness for single AS factor (concordance with single-component metric), we observe that: (a). IA-$nDCG$ is superior to other AS metrics and therefore may be the most intuitive as a metric to emphasise vertical orientation (**VS**); (b). $D\#$-$nDCG$ is superior to other AS metrics and therefore may be the most intuitive as a metric to emphasise vertical diversity (**VD**); (c). AS_{RBP} and $D\#$-$nDCG$ are the most intuitive metrics to emphasise vertical topical relevance (**IS**); (d). α-$nDCG$ is the most intuitive metric to emphasise result presentation (**RP**).

- In terms of intuitiveness to combination of factors (concordance with multiple single-component metrics), we find that: (a). AS_{RBP} is most intuitive as a metric to emphasise both vertical orientation and vertical topical relevance (**VS+IS**); (b). $D\#$-$nDCG$ is the most intuitive metric to emphasise vertical orientation, vertical topical relevance and vertical diversity (**VS+IS+VD**); (c). AS_{RBP} is the most intuitive metric to emphasise all AS components (**VS+IS+VD+RP**).

In terms of both discriminative power and intuitiveness, we demonstrate that AS-metrics (especially AS_{RBP} in this work) are the most powerful metrics to evaluate aggregated search. In addition, our work demonstrate a framework to conduct meta-evaluation for aggregated search by utilising test collection. This is relatively cheap to conduct, compared with previous work that involved with human subjects [Zhou et al., 2012c] described in Chapter 7.

Here we summarize the evaluation metric part of this thesis (Part III). In Part III of this thesis, we study another component of Cranfield paradigm, i.e. the evaluation metrics. Basically, we propose a set of aggregated search metrics that utilize the assessments available to model the user behavior and enable reliable and trustworthy evaluation of aggregated search. We start our investigations by studying evaluating one key component of aggregated search: vertical selection. We propose the risk-aware vertical selection metrics that aims to study a number of vertical selection approaches with respect to this. Secondly, in order to measure the entire aggregated search systems, we propose a general utility-effort framework to evaluate the ultimate aggregated search pages. We demonstrate the fidelity (predictive power) of the proposed metrics by correlating the metrics' score to the user preferences of aggregated search. Furthermore, in order to further compare with a suite of IR metrics for evaluating aggregated search, we meta-evaluate the reliability and intuitiveness of a variety of twelve metrics. We show that our proposed aggregated search metrics are the most reliable and intuitive metrics, compared with adapted diversity-based and traditional IR metrics.

The main conclusion of this part is that we demonstrate the feasibility to apply Cranfield Paradigm for aggregated search for reliable and trustworthy evaluation. This chapter concludes further on understanding the metric space. In the next part (Part IV), we draw conclusions from all research chapters and present our main findings and directions for future work.

Part IV

Conclusions and Discussions

9
Conclusions

In this thesis, we presented work towards enabling evaluating aggregated search using the Cranfield paradigm, which provides reproducible, reliable and trustworthy measurements for ranking aggregated search systems. Compared to traditional IR evaluation, this is particularly difficult due to the more complex user behavior on the aggregated search page that requires refined modelling. The complexity of user behavior results from the heterogeneous nature of the documents, the richer interaction that the user could operate on and the more complex presentation strategies that are operationalized. Cranfield paradigm for aggregated search consists of building a test collection and developing a set of metrics. A test collection should contain results from a set of verticals, some information needs relating to this task and relevance assessments. The metrics proposed utilized the information in the test collection in order to measure the performance of any aggregated search pages. The more complex user behaviors of aggregated search are aimed to be reflected and modeled in the metrics.

The six research chapters addressed the challenges of aggregated search Cranfield paradigm as follows. Firstly, we aim to understand better on how to make assessments for aggregated search and therefore build a reliable and reusable test collection for this task from Chapter 3 to 5 (Part II). In particular, in Chapter 3, we start our investigations by studying whether different underlying assumptions made for vertical relevance assessments affect a user's perception of the relevance of verticals. In Chapter 4, we study how to efficiently and cheaply build a test collection for aggregated search. We propose a reusing methodology to create a reliable test collection from existing test collections and available document relevance assessments. In Chapter 5, given that we collected both vertical-level (orientation) and document-level relevance assessments, we investigate whether both types of assessments somewhat correlate. The key question we answered is that we found the collection-based vertical relevance can be aligned with the user vertical intent (orientation) for evaluation purposes.

Secondly, we aim to investigate how to model the aggregated search user in a principled way in order to propose reliable, intuitive and trustworthy evaluation metrics to measure the user experience from Chapter 6 to 8 (Part III). In particular, in Chapter 6, we start our investigations by studying evaluating one key component of aggregated search: vertical selection. We propose the risk-aware vertical selection metrics that aims to study a number of vertical selection approaches with respect to both reward and risk in selecting relevant verticals. In Chapter 7, we turn to a more thorough evaluation of the entire aggregated search system, i.e. measuring the effectiveness of the ultimate aggregated search pages. We formalize the layout of the blended aggregated search page and propose a utility-effort evaluation framework to capture the user behavior. We also correlates the metric prediction with the user preferences of aggregated search pages. In Chapter 8, we further meta-evaluate a broad range of evaluation metrics for the aggregated search tasks, with respect to their reliability and intuitiveness.

Below, we provide a more detailed summary of the contributions and results of our research, and answer the research questions set out at the beginning of this thesis (Section 9.1). We conclude with an outlook on future research directions (Section 9.2).

9.1 Summaries of Main Findings

The broad question that motivates the research in this thesis is: *Can we adapt the traditional Cranfield paradigm evaluation approach (i.e. test collection based evaluation) to measure the performance of the aggregated search systems?*

The work can be splited mainly into two parts: assessments (test collection) and metrics (evaluation measures). The main objective of this thesis is to model the most essential components of the complex user behavior on aggregated search pages and incorporate them into both the assessments and evaluation metric space. We demonstrate the feasibility of this approach steps by steps as follows.

9.1.1 Aggregated Search Assessments

As *vertical* is the key component that has been introduced in aggregated search (compared to other traditional IR tasks), we start our investigations by studying whether different underlying assumptions made for vertical relevance affects a user's perception of the relevance of verticals. In Chapter 3, we formalize the different assumptions made by prior work, design and conduct several user studies in order to answer the following questions:

> **RQ 1**: Are there any differences between the assessments made by users from a pre-retrieval user-need perspective (viewing only vertical labels prior to seeing the final SERP) and the assessments made by users from a post-retrieval user perspective (viewing the vertical results in the final SERP)?
>
> **RQ 2**: When using "general web" results as a reference for making vertical relevance assessments, are these assessments able to predict the users' pairwise preference between any two verticals?
>
> **RQ 3**: Does the context (results returned from other verticals) in which vertical results are presented affect a user's perception of the relevance of the vertical of interest?
>
> **RQ 4**: Is the vertical preference information provided by a population of users able to predict the "perfect" embedding position of a vertical?

We found that both orientation (pre-retrieval user need) and topical relevance (post-retrieval topical relevance) correlates significantly with the post-retrieval search results utility. The impact of orientation is comparatively more significant (moderate) than topical relevance (low). In addition, there is an aesthetic bias to a user's perception of search results utility. We also observe that "general web" results can be served as a reference for deciding vertical relevance and it is effective from the utility-effort perspective in collection assessments. In addition, the context of other verticals has significant impact on the relevance of a vertical. We found that it is possible to employ a number of binary assessments to predict multi-grade assessments and the correlation of the derived optimal pages is significant (moderate). Using a larger number of assessments contributes to more accurate estimation of multi-grade assessments.

After analyzing how to gather vertical-level relevance assessments, in Chapter 4, we aim to build a TREC-style aggregated search test collection that collect document-level relevance assessments. This TREC-style aggregated search test collection should consist of a set of verticals, each populated by a set of items (documents), a set of topics (information needs) related to one or multiple verticals, and a set of relevance assessments between any pair of topic and item. We created a test collection by re-using a current web-based test collection and answer the questions:

> **RQ 5**: Can we reuse existing test collections to construct a test collection for aggregated search?
>
> **RQ 6**: Is the constructed test collection reliable? What is the impact of misclassification (of items into verticals) to the evaluation of systems?

We demonstrated that by identifying topics from existing test collections, a sufficient number of topics with multiple vertical intents can be collected. In addition, through simulation we have showed that aggregated search approaches can be properly evaluated even there are inherent misclassification within the verticals.

After constructing a test collection with both vertical-level and document-level relevance assessments, in Chapter 5, we turn to study whether there is some correlation between vertical-level assessments (from the pre-retrieval user intent perspective) and the one obtained by analyzing the distribution of document relevance within the vertical. We design a set of approaches to derive vertical relevance from document relevance and conduct several experiments to investigate:

> **RQ 7**: Can the vertical relevance be derived from document relevance judgments and therefore ranked similarly to the user vertical intent (orientation)?
>
> **RQ 8**: Can we appropriately threshold the derived vertical rankings and ultimately align them with the binary vertical selection decision made by the users?

We found that the alignment between collection-based relevant verticals and user vertical intents has moderate (and significant) correlation. We conclude that we might be able to use collection-based judgments as the approximate ground-truth to evaluate vertical selection (and vice versa).

9.1.2 Aggregated Search Evaluation Metrics

After addressing the above questions regarding the assessments for aggregated search that should be used to sufficiently model the user to build a test collection, we turn to the other key component for Cranfield paradigm evaluation: the evaluation metrics. Given the vertical-level assessments and document-level assessments and different stages of aggregated search, we investigate how to effectively measure the system performance.

In Chapter 6, we study how to take both the risk and reward in evaluating one key component of aggregated search: vertical selection. We use the proposed vertical selection metric to study a number of vertical selection approaches in order to answer:

> **RQ 9**: For evaluating vertical selection, rather than solely consider reward (selecting relevant verticals), can we measure the performance on maximising reward while minimising risk (selecting irrelevant verticals)?
>
> **RQ 10**: How effective and robust are existing vertical selection approaches considering the varying types of user (risk-averse and risk-seeking)?

We show that vertical selection can be measured by a proposed risk-aware VS evaluation metric that allows systems to be evaluated across a population of users, where users may have varying levels of risk (risk-averse vs. risk-seeking). Under this evaluation framework, we demonstrate that ReDDE is the most effective VS approach and CRCS(l) is the most robust VS approach.

Following the key component evaluation, we turn to a thorough evaluation of the entire aggregated search system, i.e. measuring the effectiveness of the ultimate aggregated search pages. We formalize the layout of the blended aggregated search page and propose a utility-effort evaluation framework to capture the user behavior in order to answer the following questions:

> **RQ 11**: Do users agree with each other when assessing the preference of aggregated search pairs?
>
> **RQ 12**: Can we evaluate aggregated search pages (the whole aggregated search systems) that capture both effort and utility (relevance) in a formal way? How can we utilize (combine) both vertical relevance and document relevance when evaluating aggregated search pages?
>
> **RQ 13**: Do those aggregated search metrics possess strong *predictive power*, i.e. aligning with the real user preference of aggregated search pages?
>
> **RQ 14**: Can we personalize the evaluation based on each types of user?

We show that the proposed aggregated search evaluation framework and the corresponding metrics can correlate well with the majority user preferences and that traditional IR metrics are not well suited to the task. We also showed that our metrics have the ability to tune their behaviour for pages for which personalised preference data is available.

Finally, for all different types of metrics, we aim to understand how well different metrics perform regarding both their reliability and intuitiveness:

RQ 15: How do all different suites of metrics (traditional IR, diversity IR and aggregated search) perform with respect to reliability, i.e. the ability to statistically discriminate aggregated search systems?

RQ 16: Are all different suites of metrics perform sufficiently intuitive to capture different key components of aggregated search?

In terms of both discriminative power and intuitiveness, we demonstrate that our proposed AS-metrics (especially AS_{RBP}) are the most powerful metrics to evaluate aggregated search. In addition, our work demonstrates a framework to conduct meta-evaluation for aggregated search by utilizing the test collection. This is relatively cheap to conduct, compared with the work that involved with human subjects.

9.1.3 Main Conclusion

To summarize all the contributions described above, in this thesis, we mainly demonstrate the feasibility to apply Cranfield Paradigm for aggregated search for reproducible, cheap, reliable and trustworthy evaluation.

9.2 Future Work

Due to the heterogeneous nature of information in aggregated search, numerous challenges have arisen. In this thesis, we argue that, compared with traditional homogeneous search, evaluation in the context of heterogeneous information is more challenging and requires taking into account more complex user behaviours and interactions. Specifically, this opens up many interesting and important directions for future work [Zhou et al., 2013c]. More refined evaluation approaches are required that not only model user behaviours but also adapt to how users interact with an heterogeneous information space.

9.2.1 Challenges

There are three main challenges in incorporating user behaviours within an evaluation framework for heterogeneous information access. We discuss each challenge and current research endeavours for each below.

Non-linear Traversal Browsing

Presenting heterogeneous information is more complex than the typical single ranked list (e.g. ten blue links) employed in homogeneous ranking. There are three main types of presentation designs: (1) results from the different verticals are blended into a single list (of blocks), referred to as *blended*; (2) results from each vertical are presented in a separate (e.g. horizontal paralleled) panel (tile), referred to as *non-linear blended*; and (3) vertical results can be accessed in separate tabs, referred to as *tabbed*. A combination of all three is also possible. In addition, results from different vertical search engines can be grouped together to form a coherent “bundle” for a given aspect of the query (e.g. a bundle composed of a news article along with videos and user comments as a response to a query "football match"). Finally, the results presented on the search page can contain visually salient snippets (e.g. image).

Other than blended presentation strategies investigated in this thesis, more work is required to investigate other presentation strategies. In addition, different presentation strategies and visual saliencies imply different patterns of user interaction. For example, a user could follow a non-linear traversal browsing pattern. Through eye-tracking studies [Wang et al., 2013] and search log analysis [Chen et al., 2012, Sushmita et al., 2010], recent studies have shown that in a *blended* presentation, users tend first to examine results from one vertical (vertical bias), in particular those with visual salient snippets, and results nearby. In addition, when the vertical results are not presented at the top of the search result page, users tend to scan back to re-examine previous web results either bottom-up and top-down. When presenting in a *non-linear blended* style (two parallel panels/columns), a recent eye and mouse tracking study [Navalpakkam et al., 2013] showed that users tend to firstly focus on examining top results on the first column and then jump to the right panel afterwards. For a *tabbed* presentation of vertical results, the user browsing behaviour is still poorly understood.

Diverse Search Tasks

User search tasks are more complex with heterogeneous information access than traditional homogeneous ranking. A search task's vertical orientation can affect user behaviours. Previous research [Sushmita et al., 2010] showed that the strength of a user's search task's orientation towards a particular source (vertical) type is different, and this affects user's search behaviour (for example, click-through rates). Recently, Zhou et al. [Zhou et al., 2013a] found that when assessing vertical relevance, a search task's vertical orientation is more important than the topical relevance of the retrieved results. Secondly, search task's complexity can also have a major effect on user behaviours. This is because users can access results from different verticals to accomplish their search tasks in multiple search sessions. [Arguello et al., 2012] showed that more complex search tasks require significantly more user interaction and more examination of vertical results. Finally, [Bron et al., 2013] found that user's preference for aggregated search presentation (*blended* and *tabbed*) changes during multi-session search tasks. The underlying tasks should be investigated further to understand the impact.

Coherence, Diversity and Personalization

Another important consideration when evaluating heterogeneous information access is coherence. This refers to the degree to which results from different verticals focus on a similar "sense" of the query (can they form a bundle?). Recent research [Arguello and Capra, 2012] showed that query-senses associated with the *blended* vertical results can influence user interaction with web search results. The diversity of the results is another interesting problem. It has been shown to be considerably different, and that users often have their own *personalized* vertical diversity preferences [Zhou et al., 2012c]. Finally, Santos et al. [Santos et al., 2011] showed that for an ambiguous or multi-faceted query, user's intended information need varies considerably across different verticals. These should be investigated further.

9.2.2 Avenues of Research

We need an approach that models the above mentioned user behaviours and incorporates them into system-oriented measures to evaluate heterogeneous information access. This requires two main lines of research: (1) understanding and modelling users behaviours, and (2) incorporating these into the evaluation. We elaborate on each below.

The first line of research aims to give insights on the user perspectives and provide better models of user behaviours. Although there have been studies aiming at better understanding the behaviour of users in aggregated search, the problem of evaluating heterogeneous information access is far from solved. There remains a large gap between understanding user behaviours in this context and incorporating this understanding into the evaluation measures. There has been attempts at building models of aggregated search clicks [Chen et al., 2012, Wang et al., 2013] which could be incorporated in measures, e.g. to account for search task's vertical orientation and vertical visual saliency. However, many aspects still lack investigation (e.g. coherence, diversity). We propose to follow current research endeavours and investigate models to capture user aspects, in particular those poorly accounted for in the evaluation. To achieve this, we must collect data on user behaviours for aggregated search through laboratory experiments, crowd-sourcing or accessing search engine logs.

The second line of research aims to incorporate these new models into a general evaluation framework that can accurately capture the variations in user behaviours. There are few powerful evaluation frameworks that we could use for this, for instance, TBG [Smucker and Clarke, 2012] and U-measure [Sakai and Dou, 2013] as mentioned in Section 1. Zhou et al. [Zhou et al., 2012c] also recently proposed a general evaluation framework to model utility and effort in aggregated search. In addition, we could follow [Chuklin et al., 2013] and convert obtained aggregated search click models into system-oriented evaluation. Preference-based evaluation approach is another direction that is worth of attention, for instance, [Chandar and Carterette, 2013] and [Arguello et al., 2011b].

In addition, we can focus on evaluation not only specific to aggregated search in the context of desktop environment. Firstly, various types of implicit user feedback beyond clicks have to be captured and modeled, e.g., mouse movements, page scrolls, touch gestures, etc. and ultimately incorporated into the evaluation framework. Secondly, user behavior on increasingly more popular mobile devices, with limited interaction capabilities and more user context can be modeled. The user models on such devices can be developed and plugged into the evaluation framework we proposed in order to conduct reliable aggregated search evaluation to specific devices.

This section advocates the need to incorporate user behaviours into system-oriented measures for evaluating heterogeneous information access. We listed challenges and proposed some avenues for shaping future research in this direction. A new track at TREC, FedWeb[1] [Demeester et al., 2013], is studying information access for heterogeneous information, and is the perfect forum to carry some of the research avenues discussed in this section.

[1] Federated web search: https://sites.google.com/site/trecfedweb/.

Bibliography

Cross-language evaluation forum (clef). URL `http://www.clef-campaign.org/`. (Cited on page 7.)

Nii testbeds and community for information access research (ntcir) projects. URL `http://research.nii.ac.jp/ntcir/`. (Cited on page 7.)

(Cited on page 3.)

Text retrieval conference (trec), a. URL `http://http://trec.nist.gov/`. (Cited on page 7.)

Trec web track, b. URL `http://http://plg.uwaterloo.ca/~trecweb/`. (Cited on page 9.)

Jupiter research survey, 2008. URL `http://www.iprospect.com/about/researchstudy_2008_blendedsearchresults.htm`. (Cited on page 3.)

The clueweb09 dataset., 2009. URL `http://boston.lti.cs.cmu.edu/Data/clueweb09/`. (Cited on page 16.)

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *WSDM*, pages 5–14, 2009. (Cited on pages 19, 104, 113, 116, and 129.)

Omar Alonso, Daniel E Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. In *ACM SigIR Forum*, volume 42, pages 9–15. ACM, 2008. (Cited on page 20.)

Jaime Arguello and Robert Capra. The effect of aggregated search coherence on search behavior. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1293–1302. ACM, 2012. (Cited on pages 27, 29, and 140.)

Jaime Arguello and Diaz Fernando. Vertical selection and aggregation. *Relevance Ranking and Vertical Search Engines*, 2013. (Cited on pages 5, 21, and 26.)

Jaime Arguello, Fernando Diaz, Jamie Callan, and Jean-Francois Crespo. 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, 2009. ACM. ISBN 978-1-60558-483-6. doi: http://doi.acm.org/10.1145/1571941.1571997. URL `http://doi.acm.org/10.1145/1571941.1571997`. (Cited on pages 3, 4, 7, 8, 22, 25, 26, 28, 29, 33, 35, 36, 38, 59, 65, 73, 76, 81, 84, 93, 101, and 114.)

Jaime Arguello, Fernando Diaz, and Jean-François Paiement. Vertical selection in the presence of unlabeled verticals. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 691–698. ACM, 2010. (Cited on pages 26, 28, 29, and 93.)

Jaime Arguello, Fernando Diaz, and Jamie Callan. Learning to aggregate vertical results into web search results. In *CIKM*, pages 201–210, 2011a. (Cited on pages 28 and 120.)

Jaime Arguello, Fernando Diaz, Jamie Callan, and Ben Carterette. A methodology for evaluating aggregated search results. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 141–152, Berlin, Heidelberg, 2011b. Springer-Verlag. ISBN 978-3-642-20160-8. URL `http://portal.acm.org/citation.cfm?id=1996889.1996909`. (Cited on pages 6, 20, 29, 33, 35, 36, 37, 38, 39, 40, 45, 46, 50, 51, 52, 58, 81, 93, 94, 102, 103, 104, 105, 108, 114, 119, 121, and 141.)

Jaime Arguello, Wan-Ching Wu, Diane Kelly, and Ashlee Edwards. 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*, pages 435–444. ACM, 2012. (Cited on pages 29 and 140.)

Jaime Arguello, Robert Capra, and Wan-Ching Wu. Factors affecting aggregated search coherence and search behavior. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1989–1998. ACM, 2013. (Cited on pages 27 and 29.)

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999. (Cited on page 16.)

Peter Bailey, Nick Craswell, Ryen W. White, Liwei Chen, Ashwin Satyanarayana, and Seyed M. M. Tahaghoghi. Evaluating whole-page relevance. In *SIGIR*, pages 767–768, 2010. (Cited on pages 29, 93, and 114.)

Marc Bron, Jasmijn Van Gorp, Frank Nack, Lotte Belice Baltussen, and Maarten de Rijke. 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*, pages 123–132. ACM, 2013. (Cited on pages 29 and 140.)

James P. Callan, Zhihong Lu, and W. Bruce Croft. Searching distributed collections with inference networks. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '95, pages 21–28, New York, NY, USA, 1995. ACM. ISBN 0-89791-714-6. doi: http://doi.acm.org/10.1145/215206.215328. URL `http://doi.acm.org/10.1145/215206.215328`. (Cited on pages 23, 27, 67, 84, and 101.)

Jamie Callan and Margaret Connell. Query-based sampling of text databases. *ACM Transactions on Information Systems (TOIS)*, 19(2):97–130, 2001. (Cited on pages 22 and 67.)

Ben Carterette, James Allan, and Ramesh Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 268–275. ACM, 2006. (Cited on page 58.)

Ben Carterette, Evgeniy Gabrilovich, Vanja Josifovski, and Donald Metzler. Measuring the reusability of test collections. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 231–240. ACM, 2010. (Cited on page 58.)

Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 611–620. ACM, 2011. (Cited on pages 18 and 82.)

Benjamin A Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Transactions on Information Systems (TOIS)*, 30(1): 4, 2012. (Cited on pages 111, 115, and 117.)

Praveen Chandar and Ben Carterette. Preference based evaluation measures for novelty and diversity. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 413–422. ACM, 2013. (Cited on pages 18 and 141.)

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: http://doi.acm.org/10.1145/1645953.1646033. URL `http://doi.acm.org/10.1145/1645953.1646033`. (Cited on pages 17, 93, 98, 99, and 113.)

Danqi Chen, Weizhu Chen, Haixun Wang, Zheng Chen, and Qiang Yang. Beyond ten blue links: enabling user click modeling in federated web search. In *WSDM*, pages 463–472, 2012. (Cited on pages 98, 99, 140, and 141.)

Aleksandr Chuklin, Pavel Serdyukov, and Maarten De Rijke. Click model-based information retrieval metrics. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 493–502. ACM, 2013. (Cited on pages 18 and 141.)

Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 659–666, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-164-4. doi: http://doi.acm.org/10.1145/1390334.1390446. URL `http://doi.acm.org/10.1145/1390334.1390446`. (Cited on pages 19, 58, 67, 93, 104, 112, 113, and 116.)

Charles LA Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 75–84. ACM, 2011. (Cited on pages 115 and 117.)

Charles LA Clarke, Luanne Freund, Mark D Smucker, and Emine Yilmaz. Report on the sigir 2013 workshop on modeling user behavior for information retrieval evaluation (mube 2013). In *ACM SIGIR Forum*, volume 47, pages 84–95. ACM, 2013. (Cited on page 17.)

Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306. ACM, 2002. (Cited on pages 23 and 84.)

Thomas Demeester, Dolf Trieschnigg, Dong Nguyen, and Djoerd Hiemstra. Overview of the trec 2013 federated web search track. TREC, 2013. (Cited on pages 8, 72, 73, and 141.)

Fernando Diaz. Integration of news content into web results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 182–191. ACM, 2009. (Cited on page 26.)

Fernando Diaz, Ryen White, Georg Buscher, and Dan Liebling. Robust models of mouse movement on dynamic web search results pages. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1451–1460. ACM, 2013. (Cited on page 18.)

Georges Dupret. User Models to Compare and Evaluate Web IR Metrics. In Charles Clarke, David Evans, Donna Harman, and Dianne Kelly, editors, *SIGIR 2009 Workshop on The Future of IR Evaluation*, New York, 2009. URL `http://staff.science.uva.nl/~{}kamps/ireval/papers/sue.pdf`. (Cited on page 95.)

J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. (Cited on pages 45, 75, and 103.)

Luis Gravano, Hector Garcia-Molina, and Anthony Tomasic. Precision and recall of gloss estimators for database discovery. In *Parallel and Distributed Information Systems, 1994., Proceedings of the Third International Conference on*, pages 103–106. IEEE, 1994. (Cited on pages 69 and 72.)

Donna Harman. Overview of the second text retrieval conference (trec-2). In *Proceedings of the workshop on Human Language Technology*, pages 351–357. Association for Computational Linguistics, 1994. (Cited on page 16.)

David Hawking and Paul Thomas. Server selection methods in hybrid portal search. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 75–82. ACM, 2005. (Cited on page 39.)

Dzung Hong and Luo Si. Search result diversification in resource selection for federated search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 613–622. ACM, 2013. (Cited on page 25.)

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446, October 2002. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/582415.582418. URL `http://doi.acm.org/10.1145/582415.582418`. (Cited on pages 16, 17, 96, 99, 104, 112, and 116.)

Luo Jie, Sudarshan Lamkhede, Rochit Sapra, Evans Hsu, Helen Song, and Yi Chang. 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*, pages 1195–1203. ACM, 2013. (Cited on pages 28 and 29.)

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002. (Cited on page 20.)

Thorsten Joachims et al. Evaluating retrieval performance using clickthrough data., 2003. (Cited on page 20.)

Karen S. Jones and C. J. van Rijsbergen. Report on the Need for and Provision of an ”Ideal” Information Retrieval Test Collection. British Library Research and Development Report 5266, University of Cambridge, 1975. (Cited on pages 16 and 95.)

Ioannis Kanaris and Efstathios Stamatatos. Learning to recognize webpage genres. *Information Processing & Management*, 45(5):499–512, 2009. (Cited on page 63.)

Jaana Kekäläinen and Kalervo Järvelin. Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology*, 53 (13):1120–1129, 2002. (Cited on page 72.)

Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(12):1–224, 2009. (Cited on page 20.)

Jinyoung Kim and W Bruce Croft. Building pseudo-desktop collections. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 39–40, 2009. (Cited on page 58.)

Jinyoung Kim and W Bruce Croft. Ranking using multiple document types in desktop search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 2010. (Cited on pages 4 and 5.)

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009. (Cited on pages 6, 15, 19, and 20.)

Mounia Lalmas. Aggregated search. In *Advanced Topics in Information Retrieval*, pages 109–123. Springer, 2011. (Cited on page 21.)

John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, pages 21–26, 2010. (Cited on page 21.)

Teerapong Leelanupab, Guido Zuccon, and Joemon M. Jose. A query-basis approach to parametrizing novelty-biased cumulative gain. In *Proceedings of the Third international conference on Advances in information retrieval theory*, ICTIR'11, pages 327–331, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-23317-3. URL `http://dl.acm.org/citation.cfm?id=2040317.2040360`. (Cited on page 106.)

Xiao Li, Ye-Yi Wang, and Alex Acero. Learning query intent from regularized click graphs. In *SIGIR*, pages 339–346, 2008. (Cited on pages 26, 28, 29, 93, and 114.)

Yuanhua Lv, Ariel Fuxman, and Ashok K Chandra. Evaluation of ir applications with constrained real estate. In *Advances in Information Retrieval*, pages 160–171. Springer, 2014. (Cited on page 27.)

Ilya Markov, Avi Arampatzis, and Fabio Crestani. On cori results merging. In *Advances in Information Retrieval*, pages 752–755. Springer, 2013a. (Cited on page 27.)

Ilya Markov, Leif Azzopardi, and Fabio Crestani. Reducing the uncertainty in resource selection. In *Advances in Information Retrieval*, pages 507–519. Springer, 2013b. (Cited on page 25.)

Weiyi Meng, Zonghuan Wu, Clement Yu, and Zhuogang Li. A highly scalable and effective method for metasearch. *ACM Transactions on Information Systems (TOIS)*, 19(3):310–335, 2001. (Cited on page 4.)

Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27:2:1–2:27, December 2008. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/1416950.1416952. URL `http://doi.acm.org/10.1145/1416950.1416952`. (Cited on pages 17, 98, 99, 105, and 113.)

Vanessa Murdock and Mounia Lalmas. Workshop on aggregated search. In *ACM SIGIR Forum*, volume 42, pages 80–83. ACM, 2008. (Cited on pages 4, 59, and 63.)

Vidhya Navalpakkam, LaDawn Jentzsch, Rory Sayres, Sujith Ravi, Amr Ahmed, and Alex Smola. 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*, pages 953–964. International World Wide Web Conferences Steering Committee, 2013. (Cited on pages 18 and 140.)

Dong Nguyen, Thomas Demeester, Dolf Trieschnigg, and Djoerd Hiemstra. Federated search in the wild: the combined power of over a hundred search engines. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1874–1878. ACM, 2012. (Cited on pages 120 and 121.)

Ashok Kumar Ponnuswami, Kumaresh Pattabiraman, Desmond Brand, and Tapas Kanungo. 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*, pages 67–76. ACM, 2011a. (Cited on pages 7 and 29.)

Ashok Kumar Ponnuswami, Kumaresh Pattabiraman, Qiang Wu, Ran Gilad-Bachrach, and Tapas Kanungo. On composition of a federated web search result page: using online users to provide pairwise preference for heterogeneous verticals. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 715–724, New York, NY, USA, 2011b. ACM. ISBN 978-1-4503-0493-1. doi: http://doi.acm.org/10.1145/1935826.1935922. URL `http://doi.acm.org/10.1145/1935826.1935922`. (Cited on pages 7, 8, 28, 29, 33, 35, 36, 37, 39, 40, 54, 58, 82, 93, 94, and 114.)

Allison L Powell and James C French. Comparing the performance of collection selection algorithms. *ACM Transactions on Information Systems (TOIS)*, 21(4):412–456, 2003. (Cited on pages 69 and 72.)

Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 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. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458092. URL `http://doi.acm.org/10.1145/1458082.1458092`. (Cited on pages 6, 15, and 20.)

Ian Ruthven and Diane Kelly. *Interactive information seeking, behaviour and retrieval.* Facet Publ., 2013. (Cited on page 15.)

Tetsuya Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 525–532. ACM, 2006. (Cited on pages 111 and 117.)

Tetsuya Sakai. Evaluation with informational and navigational intents. In *Proceedings of the 21st international conference on World Wide Web*, pages 499–508. ACM, 2012. (Cited on pages 111, 115, 117, and 118.)

Tetsuya Sakai and Zhicheng Dou. Summaries, ranked retrieval and sessions: a unified framework for information access evaluation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 473–482. ACM, 2013. (Cited on pages 18 and 141.)

Tetsuya Sakai and Ruihua Song. Evaluating diversified search results using per-intent graded relevance. In *SIGIR*, pages 1043–1052, 2011. (Cited on pages 19, 93, 104, 113, 115, and 116.)

Tetsuya Sakai and Ruihua Song. Diversified search evaluation: Lessons from the ntcir-9 intent task. *Information retrieval*, 16(4):504–529, 2013. (Cited on page 115.)

Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010. (Cited on pages 6, 7, 15, and 16.)

Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up? In *SIGIR*, pages 555–562, 2010. (Cited on pages 7, 16, 46, 72, 103, 104, and 115.)

Marina Santini. *Automatic identification of genre in web pages*. PhD thesis, University of Brighton, 2007. (Cited on pages 61 and 63.)

Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. Aggregated search result diversification. In *Proceedings of the 3rd International Conference on the Theory of Information Retrieval*, pages 250–261, Bertinoro, Italy, 2011. Springer. (Cited on pages 30, 94, and 140.)

Markus Schulze. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Social Choice and Welfare*, 36 (2):267–303, 2011. (Cited on page 38.)

Jangwon Seo and W Bruce Croft. Blog site search using resource selection. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1053–1062. ACM, 2008. (Cited on pages 25 and 84.)

Milad Shokouhi. 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, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-71494-1. URL `http://dl.acm.org/citation.cfm?id=1763653.1763674`. (Cited on pages 25, 67, 84, and 120.)

Milad Shokouhi and Luo Si. Federated search. *Foundations and Trends in Information Retrieval*, 5(1):1–102, 2011. (Cited on pages 4, 5, 21, 22, 29, 37, 39, and 58.)

Milad Shokouhi and Justin Zobel. Robust result merging using sample-based score estimates. *ACM Transactions on Information Systems (TOIS)*, 27(3):14, 2009. (Cited on pages 27 and 28.)

Milad Shokouhi, Justin Zobel, Saied Tahaghoghi, and Falk Scholer. Using query logs to establish vocabularies in distributed information retrieval. *Information processing & management*, 43(1):169–180, 2007. (Cited on page 22.)

Luo Si and Jamie Callan. Using sampled data and regression to merge search engine results. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26. ACM, 2002. (Cited on page 24.)

Luo Si and Jamie Callan. 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. ACM. ISBN 1-58113-646-3. doi: http://doi.acm.org/10.1145/

860435.860490. URL http://doi.acm.org/10.1145/860435.860490. (Cited on pages 22, 24, 59, 67, 84, 101, and 120.)

Luo Si and Jamie Callan. A semisupervised learning method to merge search engine results. *ACM Transactions on Information Systems (TOIS)*, 21(4):457–491, 2003b. (Cited on page 27.)

Luo Si and Jamie Callan. Modeling search engine effectiveness for federated search. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 83–90. ACM, 2005. (Cited on page 26.)

Mark D Smucker and Charles LA Clarke. Time-based calibration of effectiveness measures. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 95–104. ACM, 2012. (Cited on pages 17 and 141.)

Shanu Sushmita, Hideo Joho, Mounia Lalmas, and Robert Villa. Factors affecting click-through behavior in aggregated search interfaces. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 519–528. ACM, 2010. (Cited on pages 6, 29, 38, 69, and 140.)

Paul Thomas. Server characterisation and selection for personal metasearch. In *SIGIR Forum*, volume 42, pages 108–109. Citeseer, 2008. (Cited on page 25.)

Andrew Turpin and Falk Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 11–18. ACM, 2006. (Cited on page 6.)

Ellen M Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36(5):697–716, 2000. (Cited on page 45.)

Ellen M. Voorhees. Overview of the trec 2003 question answering track. In *TREC*, pages 54–68, 2003. (Cited on page 104.)

Chao Wang, Yiqun Liu, Min Zhang, Shaoping Ma, Meihong Zheng, Jing Qian, and Kuo Zhang. Incorporating vertical results into search click models. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 503–512. ACM, 2013. (Cited on pages 69, 140, and 141.)

Lidan Wang, Jimmy Lin, and Donald Metzler. Learning to efficiently rank. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 138–145. ACM, 2010. (Cited on page 18.)

Xiao-Bing Xue, Zhi-Hua Zhou, and Zhongfei (Mark) Zhang. Improving web search using image snippets. *ACM Trans. Internet Technol.*, 8:21:1–21:28, October 2008. ISSN 1533-5399. doi: http://doi.acm.org/10.1145/1391949.1391955. URL http://doi.acm.org/10.1145/1391949.1391955. (Cited on page 98.)

Cheng Xiang Zhai, William W Cohen, and John Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 10–17. ACM, 2003. (Cited on pages 19 and 113.)

Ke Zhou, Ronan Cummins, Mounia Lalmas, and Joemon Jose. Evaluating large-scale distributed vertical search. In *Proceedings of the 9th workshop on Large-scale and distributed informational retrieval*, LSDS-IR '11, pages 9–14, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0959-2. doi: 10.1145/2064730.2064735. URL `http://doi.acm.org/10.1145/2064730.2064735`. (Cited on pages 42, 86, 100, 119, and 121.)

Ke Zhou, Ronan Cummins, Martin Halvey, Mounia Lalmas, and Joemon M. Jose. Assessing and predicting vertical intent for web queries. In *ECIR*, pages 499–502, 2012a. (Cited on pages 33, 35, 36, 37, 38, 39, 40, 42, 73, 75, 82, 86, 93, 95, 100, 103, 108, 109, 114, 116, and 120.)

Ke Zhou, Ronan Cummins, Mounia Lalmas, and Joemon M Jose. Evaluating reward and risk for vertical selection. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2631–2634. ACM, 2012b. (Cited on pages 36, 52, 75, and 114.)

Ke Zhou, Ronan Cummins, Mounia Lalmas, and Joemon M Jose. Evaluating aggregated search pages. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 115–124. ACM, 2012c. (Cited on pages 36, 46, 112, 113, 114, 115, 116, 122, 129, 131, 140, and 141.)

Ke Zhou, Ronan Cummins, Mounia Lalmas, and Joemon M Jose. Which vertical search engines are relevant? In *Proceedings of the 22nd international conference on World Wide Web*, pages 1557–1568. International World Wide Web Conferences Steering Committee, 2013a. (Cited on pages 69, 73, and 140.)

Ke Zhou, Mounia Lalmas, Tetsuya Sakai, Ronan Cummins, and Joemon M Jose. On the reliability and intuitiveness of aggregated search metrics. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 689–698. ACM, 2013b. (Cited on page 112.)

Ke Zhou, Tetsuya Sakai, Mounia Lalmas, Zhicheng Dou, and Joemon M Jose. Evaluating heterogeneous information access (position paper). *Clarke et al.[19]*, pages 19–20, 2013c. (Cited on page 139.)

Ke Zhou, T. Demeester, D. Nguyen, D. Hiemstra, and D.. Trieschnigg. Aligning vertical collection relevance with user intent. In *Proceedings of the 23nd ACM international conference on Conference on information & knowledge management*. ACM, 2014. (Cited on page 69.)

Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 307–314. ACM, 1998. (Cited on pages 16 and 65.)

Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Emine Yilmaz, Joemon M. Jose, and Leif Azzopardi. Crowdsourcing interactions: Using crowdsourcing for evaluating interactive information retrieval systems. *Inf. Retr.*, 16(2):267–305, April 2013. ISSN 1386-4564. doi: 10.1007/s10791-012-9206-z. URL `http://dx.doi.org/10.1007/s10791-012-9206-z`. (Cited on page 20.)