Which Vertical Search Engines are Relevant?

Understanding Vertical Relevance Assessments for Web Queries

Ke Zhou
University of Glasgow
Glasgow, United Kingdom
zhouke@dcs.gla.ac.uk

Ronan Cummins
University of Greenwich
London, United Kingdom
r.p.cummins@gre.ac.uk

Mounia Lalmas
Yahoo! Labs
Barcelona, Spain
mounia@acm.org

Joemon M. Jose
University of Glasgow
Glasgow, United Kingdom
jj@dcs.gla.ac.uk

ABSTRACT

Aggregating search results from a variety of heterogeneous sources, so-called verticals, such as news, image and video, into a single interface is a popular paradigm in web search. Current approaches that evaluate the effectiveness of aggregated search systems are based on rewarding systems that return highly *relevant* verticals for a given query, where this *relevance* is assessed under different assumptions. It is difficult to evaluate or compare those systems without fully understanding the relationship between those underlying assumptions. To address this, we present a formal analysis and a set of extensive user studies to investigate the effects of various assumptions made for assessing query vertical *relevance*. A total of more than 20,000 assessments on 44 search tasks across 11 verticals are collected through Amazon Mechanical Turk and subsequently analysed. Our results 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.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]
Keywords: aggregated search, vertical selection, evaluation, user study, relevance assessment

1. INTRODUCTION

Aggregated search originated from federated search [12], which studies the simultaneous retrieval of information from various separate indexes. With the increasing amount of different types of online media (e.g. images, news, video), it is becoming popular for web search engines to present results from a set of specific verticals dispersed throughout the standard ‘general web’ results, for example, adding image results to the ten blue links for the query “yoga poses”. Retrieving and integrating these various information sources into one interface is called *aggregated search* (AS) and has become the de-facto search paradigm in commercial search engines.

A key component of AS is *vertical selection* (VS): *selecting relevant verticals (if any) from which items will be selected to appear on the search result page (SERP) alongside the ‘general web’ search results for a given query.* This has been studied in several work [3, 4, 8, 1] and various solutions have been proposed. Evaluating VS approaches [3, 16, 2, 9, 18, 17] is a challenging problem. Much of the research to date assumes that zero to many verticals are pertinent to a particular query, and aims to compare the quality of a set of selected verticals against an annotated set[1]. These type of studies all assume that the annotation set is available and is obtained by either, explicitly collecting labels from assessors [3, 16, 2, 9, 18], or implicitly, by deriving them from user interaction information [9, 5]. Despite the relative success of these evaluation methodologies, the definition of the *relevance* of a vertical, given a query, remains unclear. Different work makes different assumptions when obtaining the assessments for *relevant* verticals across a set of queries.

In this paper, 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.

Firstly, 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). 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 informa-

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

[1] A set of verticals annotated by a user (or users) for a given query.

tion already presented in a *'video'* result). Finally, when AS 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. Understanding these underlying assumptions when assessing the relevance of verticals is important. 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.

Although existing work collects assessments (using different processes and assumptions) for evaluating VS, to our knowledge, no work has tried to comprehensively understand and compare those assessment processes and assumptions. This is the focus of this paper. Specifically, we employ all of the various strategies present in the literature, to collect vertical relevance assessments, and investigate three main research questions (RQ):

- **(RQ1)** 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)?
- **(RQ2)** 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? 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?
- **(RQ3)** Is the preference information provided by a population of users able to predict the "perfect" embedding position of a vertical?

To answer these three research questions, we conducted a set of large-scale user studies using the crowd-sourcing platform Amazon Mechanical Turk. In Section 2, we formally outline the problem of vertical selection assessment. Section 3 outlines our experimental design, whereas in Section 4 we present and analyse our results. We conclude the paper in Section 5 by summarising all the findings, discussing the implications of our results and pointing out limitations.

2. 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.

2.1 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 [9], 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 [3, 16], 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 [9, 2]. By assigning diminishing weight according to the embedding position[2], each weight $i_k \in I_t$ of vertical v_k is represented by the corresponding assigned weight of v_k's 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 [9, 16, 3, 2, 18, 17] 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 [3, 16]. Recently, risk has been considered and incorporated into risk-aware VS metrics [18]. 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 [2]. The further the distance from the perfect page, the worse the performance of the system that generated that SERP P.

2.2 Making Preference Assessments

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

2.2.1 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

[2] 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).

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 work [9, 16, 3] follows 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's work [2] 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 [2]).

2.2.2 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 depends 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 [3, 11], 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 [11, 2] that the visually attractive nature of image results tends to increase users engagement on a SERP, compared to those that do not contain images.

2.3 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 [16, 3] 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 [2, 10] 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.

2.4 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 [12, 7] (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. [16] 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. [2], 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 [16], 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 [9, 2], a vertical relevance is assessed by presenting the SERP with the web results and vertical results separately. In Ponnuswami et al. [9], 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 [2] and [9] is that, instead of voting across all verticals, the '*general web*' retrieved results are used as an anchor to determine the vertical importance.

Table 1: **Summary of Vertical Relevance Assumptions Made in Previous Works.**

Work	Relevance Dependency		Influencing Factors			Assessment		# Assessors
	Inter-dependent	Anchor-based	Result Quality	Orientation	Aesthetic	Binary	Graded	
Federated Search [12]	✓		✓			✓		1
Zhou et al. [16]		✓		✓			✓	4
Ponnuswami et al. [9]		✓	✓	✓	✓		✓	1
Arguello et al. [2]	✓		✓	✓	✓	✓		4

2.5 Summary of Aims

We are interested in answering three research questions (RQ1 to RQ3). Given the more formal treatment of the task of aggregated search described in this section, these research questions can be stated as follows:

- **RQ1** deals with comparing the user perspective ([16] and [9] (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)?
- **RQ2** is concerned with comparing the anchor-based approach with an inter-dependent approach ([9] (binary assessment variant) and [2]) 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.
- **RQ3** 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. EXPERIMENTAL DESIGN

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

3.1 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.

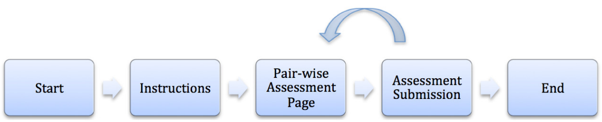

Figure 1: Flow Diagram Description of Experimental Protocol for Studies 1 to 3.

3.1.1 *Protocol*

The three studies follow a similar protocol shown in Figure 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 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[3]. 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.

3.1.2 *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")[4]. In this paper, 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 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 2.

Search tasks were chosen to have a varying number and type of relevant verticals. From a preliminary study [16, 15], 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

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

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

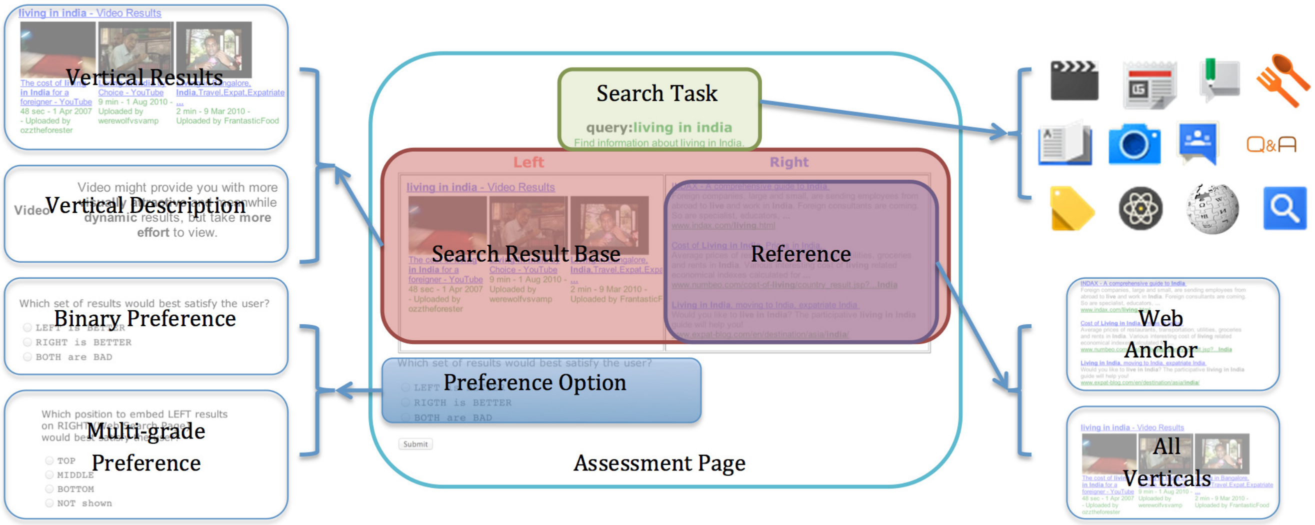
Vertical Results
living in india - Video Results
Vertical Description
Video might provide you with more dynamic results, but take more effort to view.
Binary Preference
Which set of results would best satisfy the user?
LEFT is BETTER
RIGHT is BETTER
BOTH are BAD
Multi-grade Preference
Which position to embed LEFT results on RIGHT would best satisfy the user?
TOP
MIDDLE
BOTTOM
NOT shown
Search Task
query:living in india
Find information about living in India.
Left
Right
Search Result Base
Reference
Preference Option
Which set of results would best satisfy the user?
LEFT is BETTER
RIGTH is BETTER
BOTH are BAD
Submit
Assessment Page
Q&A
Web Anchor
All Verticals

Figure 2: Various Components for Manipulations on Assessment Page of Studies 1 to 3

Table 2: Verticals Used in this Paper.

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	

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.

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. 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.

3.1.3 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 **RQ1** 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 **RQ2** 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 **RQ3** 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 indepen-

Table 3: Example Search Tasks.

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

dent 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 [6]) 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 [14]) to measure inter-assessor agreement on absolute assessments (query-document topical relevance assessment), we used Kappa measure, as prevalently used in previous work [2]. 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_{jk} 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.

3.1.4 Crowd-sourcing Data Collection

Our preference assessment data is collected over the Amazon Mechanical Turk crowd-sourcing 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 [13], “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 assesement. “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 careless 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, and 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 [17, 2]. As we have two different “trap” approaches and a large percentage of assessments are “traps”[5], we believe that our methodology was able to filter out large percentage of careless assessors.

3.2 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

[5] 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”.

an ASP and ask the assessor to provide a binary preference ("left is better" and "right is better"). To avoid overburdening assessors, we also include an option ("both are bad") that captures the scenario where a user is confused due to, for example, poor quality of both 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 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 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?).

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).

3.3 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 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).

3.4 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 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).

4. 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[6]. We denote the significance by ▲ (with $p < 0.05$).

4.1 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 pairwise 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

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

Table 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	Answer	Shopping	Discussion	Scholar	Wiki	Average
Orientation	0.547▲	0.654▲	0.864▲	0.524▲	0.516▲	0.385▲	0.563▲	0.610▲	0.305▲	0.450▲	0.404▲	0.529
Topical Relevance	0.092	0.205▲	0.637▲	0.301▲	0.187▲	0.429▲	0.354▲	0.264▲	0.571▲	0.393▲	0.484▲	0.356

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 correlation of the inter-assessor agreement (K_F) 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 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.[7] 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.

Table 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%

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 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 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.

4.2 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 [2]. 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.

[7] 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 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 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 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 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).

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[8] 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.

Table 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%

4.3 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 [2].

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[9] (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.[10]

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 [2]) 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.[11] From a closer examination, we found that this might result from each assessors' unique preference of verticals and their risk-level [18] (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 oth-

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

[9] 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.

[10] 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 $2.75 \in [2, 3)$).

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

Table 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

ers 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 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 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.

5. CONCLUSIONS AND DISCUSSIONS

Our objective was 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 [9].

Plans for future work include the following: 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 the assessment depends solely on snippet, we should examined the impact of this further.

Acknowledgments This work was partially supported by the EU LiMoSINe project (288024).

6. REFERENCES

[1] J. Arguello, F. Diaz, and J. Callan. Learning to aggregate vertical results into web search results. In CIKM 2011, pages 201-210.

[2] J. Arguello, F. Diaz, J. Callan, and B. Carterette. A methodology for evaluating aggregated search results. In ECIR 2011: pages 141-152.

[3] J. Arguello, F. Diaz, J. Callan, and J. Crespo. Sources of evidence for vertical selection. In SIGIR 2009: pages 315-322.

[4] J. Arguello, F. Diaz, and J. Paiement. Vertical selection in the presence of unlabeled verticals. In SIGIR 2010: pages 691-698.

[5] F. Diaz. Integration of news content into web results. In WSDM 2009: pages 182-191.

[6] J. Fleiss. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5): pages 378-382, 1971.

[7] D. Hawking and P. Thomas. Server selection methods in hybrid portal search. In SIGIR 2005: pages 75-82.

[8] X. Li, Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In SIGIR 2008: pages 339-346.

[9] A. Ponnuswami, K. Pattabiraman, Q. Wu, R. Gilad-Bachrach, and T. Kanungo. On composition of a federated web search result page: using online users to provide pairwise preference for heterogeneous verticals. In WSDM 2011: pages 715-724.

[10] M. Schulze. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. In Social Choice and Welfare, 2010.

[11] S. Sushmita, H. Joho, M. Lalmas, and R. Villa. Factors affecting click-through behavior in aggregated search interfaces. In CIKM 2010: pages 519-528.

[12] M. Shokouhi, and L. Si. Federated Search. Foundations and Trends in Information Retrieval (FTIR) 5(1): pages 1-102, 2011.

[13] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In SIGIR 2010: pages 555-562.

[14] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In Information Process Management. 36(5): pages 697-716, 2000.

[15] K. Zhou, R. Cummins, M. Lalmas, and J.M. Jose. Evaluating large-scale distributed vertical search. In CIKM Workshop LSDS-IR 2011.

[16] K. Zhou, R. Cummins, M. Halvey, M. Lalmas, and J. M. Jose. Assessing and predicting vertical intent for web queries. In ECIR 2012: pages 499-502.

[17] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating aggregated search pages. In SIGIR 2012: pages 115-124.

[18] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating reward and risk for vertical selection. In CIKM 2012: pages 2631-2634.