# Evaluating Aggregated Search Pages

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

Ronan Cummins
National University of Ireland
Galway, Ireland
ronan.cummins@nuigalway.ie

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

Joemon M. Jose
University of Glasgow
Glasgow, United Kingdom
joemon.jose@glasgow.ac.uk

## ABSTRACT

Aggregating search results from a variety of heterogeneous sources or *verticals* such as news, image and video into a single interface is a popular paradigm in web search. 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. This paper proposes a general framework for evaluating the quality of aggregated search pages. We evaluate our approach by collecting annotated user preferences over a set of aggregated search pages for 56 topics and 12 verticals. 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.

## Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval

## General Terms

Measurement, Experimentation

## Keywords

aggregated search, evaluation, performance metric, diversity

## 1. INTRODUCTION

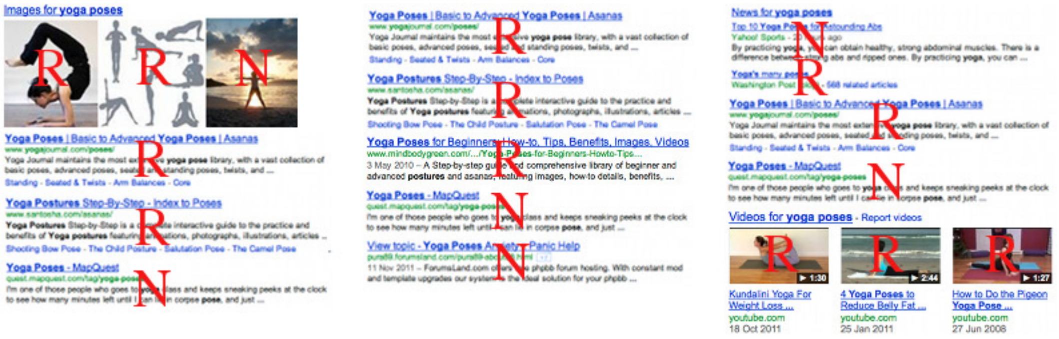
With the emergence of various vertical search engines dedicated to certain media types and genres, such as news, image, video, it is becoming popular to present results from a set of specific verticals dispersed throughout the standard "general web" results, for example by adding image results to the ten blue links for the query "pictures of flowers". This new search paradigm is often known as *aggregated search* [4]. The three main challenges that arise in realising such systems are vertical selection (**VS**), item selection (**IS**), and result presentation (**RP**). Vertical selection deals with deciding which verticals are implicitly intended by a query. Item selection deals with selecting a subset of items from each vertical to present on the aggregated page. Result presentation deals with organising and embedding the various types of items on the result page. The most common presentation strategy for aggregated search is to merge the results into one ranked list of so-called *blocks*, and is now the 'de facto' standard in many search engines.

Although various approaches exist for selecting relevant verticals [4, 5] and for optimising aggregated search pages [2, 17], evaluating the quality of aggregated search pages is still a challenge. 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 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.

In this paper, we propose a general framework for instantiating metrics that can evaluate the quality of aggregated search pages in terms of both *reward* and *effort*. Specifically, we develop an approach that uses both topical-relevance and vertical-orientation information to derive the utility of any

*SIGIR'12,* August 12–16, 2012, Portland, Oregon, USA.

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


Images for yoga poses
R
R
N
Yoga Poses | Basic to Advanced Yoga Poses | Asanas
R
Yoga Postures Step-By-Step - Index to Poses
R
Yoga Poses - MapQuest
N
Yoga Poses | Basic to Advanced Yoga Poses | Asanas
R
Yoga Postures Step-By-Step - Index to Poses
R
Yoga Poses for Beginners: How-to, Tips, Benefits, Images, Videos
R
Yoga Poses - MapQuest
N
View topic - Yoga Poses Anxiety - Panic Help
N
News for yoga poses
Top 10 Yoga Poses for Astounding Abs
N
Yoga's many poses
R
Yoga Poses | Basic to Advanced Yoga Poses | Asanas
R
Yoga Poses - MapQuest
N
Videos for yoga poses - Report videos
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 1: Three Examplar Aggregated Search Pages for the Query "yoga poses".**

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.

This paper makes several contributions: (i) we propose a framework for the evaluation of aggregated search pages that capture both effort and reward in a formal way; (ii) we outline a novel approach for simulating aggregated search pages and collect a large set of user preferences over page pairs; (iii) we demonstrate the effectiveness of the metrics derived from our framework by comparing them with several existing IR metrics; and (iv) we show that our metrics can be personalised for each user and, therefore, can be further improved using training data.

Related work is reviewed in Section 2. In Section 3, we formally outline the problem of aggregated search evaluation and list the assumptions made in this work. In Section 4, 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 5. Subsequently in Section 6, 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 and future work are discussed in Section 7.

## 2. RELATED WORK

Various works evaluating one component of an aggregated search system in isolation exist. Vertical selection in aggregated search has been studied in [4, 5, 15, 26]. 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 [4, 5, 26] or derived from user interaction data [15]. The annotation can be binary [4, 5] or graded [26]. 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 [26], as this allows for a more refined evaluation scheme.

Recent attempts to evaluate the utility of the whole aggregated search page [3, 17] 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, [17] 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 [6] 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 [3] 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 [8, 10, 18], 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 [20] have proposed an aggregated search metric that captures both vertical diversity and topical diversity. It can be noted that the framework developed in this paper can also be extended to incorporate topical diversity, but due to space limitations, we will leave this as future work.

# 3. PROBLEM FORMULATION

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

## 3.1 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 [3, 17] 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.

## 3.2 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 $qrel(I_{ij}|q)$.
- The user specific vertical-orientation [26], 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 [26] 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 [13] 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.

## 3.3 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.

# 4. 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 [11], 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

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

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

the expected gain of reading a page divided by the expected effort spent:

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

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 $nDCG$ [12]. We normalise the score of the utility of page $P$ by that of the ideal page $P_{ideal}$:

$$nUtil(P) = \frac{Util(P)}{Util(P_{ideal})} \qquad (2)$$

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.

## 4.1 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 $B_i$:

$$Orient(B_i, \alpha) = g(orient(V_j|W, q), \alpha) \qquad (3)$$

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 $\alpha$. 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:

$$g(x, \alpha) = \frac{1}{1 + \alpha^{-log_{10}(x/(1-x))}} \qquad (4)$$

A graph of the function $g(x, \alpha)$ is shown in Figure 2. This function controls how much the gain increases as the vertical-orientation level increases. When $\alpha$ 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 $\alpha$ equals to 10, the reward is exactly the same as the vertical orientation $orient(V_j|W, q)$.

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

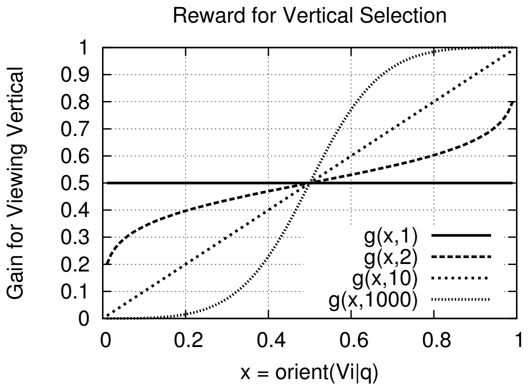


**Figure 2: Function $g()$ for Controlling Reward on Orientation with Various Parameter $\alpha$.**

$$Topic(B_i) = \sum_{k=1}^{|B_i|} qrel(I_{ik}|q) \qquad (5)$$

$|B_i|$ 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 gain based on the above two criteria:

$$G(B_i) = Orient(B_i, \alpha) \cdot Topic(B_i) \qquad (6)$$

where $\alpha$ 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.

## 4.2 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:

$$E(B_i) = \sum_{k=1}^{|B_i|} E(I_{ik}) \qquad (7)$$

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 [23], 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

**Table 1: Effort of Reading each Category.**

| Snippet Category | "image" | "text" | "video" |
|---|---|---|---|
| Effort | 1 | 3 | 6 |

time to assess as a text[4] (14 seconds). Therefore, the relative effort taken to examine each snippet type is shown in Table 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.

## 4.3 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 [8, 9, 16] that aim to predict the probability with which a user will examine an item. Position models [16] use only the position of the item in a result set. The cascade model [8] 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 [9] 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** [12] and **RBP** [16], the cascade model **ERR** [8], and the attention model **ATT** [9].

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**, $\beta_{dist}$ 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 $\zeta$, instead of assuming that $\zeta$ is a position-specific parameter, we assume that $\zeta$ 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 $\zeta$ is obtained by exploring the optimal setting in a development set.

## 4.4 Normalisation Using the Ideal Page

A summary of the non-normalised utility metrics that can be instantiated in our framework are listed in Table 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 5.2, where the simulation of aggregated page pairs is discussed.

## 4.5 Personalised Utility Metrics

Previous research [26] 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 $\lambda_i$:

$$I\_Util(P, \lambda_i) = (1 - \lambda_i) \cdot nUtil(P) + \lambda_i \cdot vRecall(P) \quad (8)$$

where $\lambda_i$ 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 $\lambda_i$ is, the more the user prefers a page with items originating from different verticals (high vertical diversity).

# 5. 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.

## 5.1 Data

We use an aggregated search test collection [25] created by reusing the existing web collection ClueWeb09. This test collection consists of a number of verticals (listed in Table 3), 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 Million-Query Track). The vertical-orientation information of each topic [26] 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?").

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 4.

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

Table 2: Summarisation of Utility Metrics for Aggregated Search.

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

Table 4: Distribution of Number of Selected Topics Assigned to Various Highly Oriented Verticals.

| Verticals | Image | Video | Recipe | News | Book | Blog | Ans | Shop | Disc | Schol | Wiki | Web-only | Total Qrys |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Topic Num | 4 | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 5 | 3 | 12 | 10 | 56 |

Table 3: Verticals Used in this Paper.

| 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 | |

## 5.2 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 [21] and CORI [7] 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 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 $nUtil$.

## 5.3 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 [3], 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

[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 [4].

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

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.

## 5.4 Collecting Pairwise Preference Assessments

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. 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 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 × 48 × 4). Following [19], 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.

## 5.5 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 want to answer the following question: **RQ1** Do users agree with each other when assessing aggregated search pairs? Therefore, we measured annotator agreement of preferences of aggregated page pairs using Fleiss' Kappa [24] (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 5.

**Table 5: 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 |

We observe that assessor agreement on presentation-pairs was $K_F = 0.241$, which is considered *fair* agreement [24]. This result is similar to previous research [3, 26], which reaffirm that evaluating aggregated 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 [3], 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.

# 6. EVALUATION

We investigate the fidelity[8] [22] of the proposed metrics. We leave an investigation on the reliability of the metric (discriminative power [18]) for future work. We aim to answer the following questions:

1. **RQ2** With standard parameter settings, are the standard diversity metrics suitable for aggregated search, and do our proposed metrics accurately predict user preferences for aggregated search pages?
2. **RQ3** Can we learn personalised parameters from historical data and, subsequently, provide a higher agreement with the user preferences?

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* [12], *P*@10), and diversity-aware metrics ($\alpha$-*nDCG* [10], *D*-*nDCG* [18], *D#*-*nDCG* [18], *IA*-*nDCG* [1]) 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.

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

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

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].

## 6.1 Standard Parameter Settings

To answer **RQ2**, 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 ($AS_{DCG}$, $nDCG$, $\alpha$-$nDCG$, $D$-$nDCG$ and $D\#$-$nDCG$). We set the $\alpha$ parameter in $\alpha$-$nDCG$ to 0.5 and $\gamma$ 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 [16]. 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 $\zeta$ 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.

### 6.1.1 Comparison of Metrics

We present results for a majority preference of 3/4 or greater, or 4/4, in Table 6. 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 [3].

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 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, $\alpha$-$nDCG$ performs significantly worse than the proposed metrics. This is because $\alpha$-$nDCG$ 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 [14] has suggested that $\alpha$ 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, $\alpha$-$nDCG$ 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.

### 6.1.2 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[10]. We observe that the metric agreements are com-

[9] We also used the sign test [3]. For all page pairs with majority of preference, our proposed metrics performed significantly better than random. Since we are interested in comparing metrics, we do not report the sign test outcomes.

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

**Table 6: Metric Agreements with Various User's Majority Preference: Proposed Metric vs. Baseline Metrics.**

| majority preference | bins | $AS_{DCG}$ | $D$-$nDCG$ | $D\#$-$nDCG$ | $IA$-$nDCG$ | $\alpha$-$nDCG$ | $nDCG$ | $P@10$ |
|---|---|---|---|---|---|---|---|---|
| 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: 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% |

paratively 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 $AS_{ATT}$.

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 $\zeta$ 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) [3]. 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.

### 6.1.3 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 [26] 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.

## 6.2 Learning for Metrics

We can improve the performance of our metrics by learning suitable parameter settings using training data, thus addressing the research question **RQ3**. We only use $AS_{RBP}$ as an example. We recall that $AS_{RBP}$ has three parameters: $\alpha$ that controls the degree to which vertical orientation is rewarded; $\beta$ that controls the user browsing behavior in terms of user persistence; and $\lambda_i$ 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 $\alpha$ and $\beta$ independently of $\lambda_i$. 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: $\alpha$ (from 1 to 100) and $\beta$ (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 8. 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 $\alpha$ and $\beta$ 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 $\alpha$ and $\beta$ and learn personalised

**Table 8: Learned $AS_{RBP}$ Metric Agreements with User's Majority Preferences for All Page Pairs.**

| Parameter | Standard | Optimal $\alpha$ and $\beta$ | Optimal $\alpha$, $\beta$ and $\lambda_i$ |
|---|---|---|---|
| Agreement | 67.7% | 72.6%▲ | 75.9%▲ |

user preference parameters for diversity ($\lambda_i$). 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 8. The optimal setting for $\lambda_i$ varies from 0.15 to 0.4 among different assessors whereas the average optimal setting for those assessors is 0.23. Similar to [26], 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.

# 7. CONCLUSIONS AND FUTURE WORK

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.

Future work will involve extending and setting several parameters of the metrics so that they more closely correlate with user preferences for the sets of page pairs. In particular, when query-log data is available, we can further extend the framework by proposing new user browsing models for aggregated search and investigate their values. We will also devise new approaches to utilise the available implicit user feedback data to better estimate the parameters. Another interesting challenge will be to compare the effectiveness and weakness of existing diversity-aware metrics for evaluating aggregated search, and study the ability of our framework to generalise them.

**Acknowledgments** This work was supported partially by both the EU LiMoSINe project (288024) and by the Irish Research Council's INSPIRE (FP7 COFUND) programme. Any opinions, findings, and recommendations expressed in this paper are the authors' and do not necessarily reflect those of the sponsors.

# 8. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. *WSDM*, 2009.

[2] J. Arguello, F. Diaz, and J. Callan. Learning to aggregate vertical results into web search results. *CIKM*, 2011.

[3] J. Arguello, F. Diaz, J. Callan, and B. Carterette. A methodology for evaluating aggregated search results. *ECIR*, 2011.

[4] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. *SIGIR*, 2009.

[5] J. Arguello, F. Diaz, and J.-F. Paiement. Vertical selection in the presence of unlabeled verticals. In *SIGIR*, 2010.

[6] P. Bailey, N. Craswell, R. W. White, L. Chen, A. Satyanarayana, and S. M. M. Tahaghoghi. Evaluating whole-page relevance. *SIGIR*, 2010.

[7] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. *SIGIR*, 1995.

[8] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. *CIKM*, 2009.

[9] D. Chen, W. Chen, H. Wang, Z. Chen, and Q. Yang. Beyond ten blue links: enabling user click modeling in federated web search. *WSDM*, 2012.

[10] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. *SIGIR*, 2008.

[11] G. Dupret. User Models to Compare and Evaluate Web IR Metrics. In *SIGIR 2009 Workshop on The Future of IR Evaluation*, 2009.

[12] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *TOIS*, 2002.

[13] K. S. Jones and C. J. Rijsbergen. Report on the need for and the provision of an 'ideal' information retrieval test collection. British Library Research and Development Report No. 5266, 1975.

[14] T. Leelanupab, G. Zuccon, and J. M. Jose. A query-basis approach to parametrizing novelty-biased cumulative gain. In *ICTIR*, 2011.

[15] X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *SIGIR*, 2008.

[16] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *TOIS*, 2008.

[17] A. K. 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. *WSDM*, 2011.

[18] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. *SIGIR*, 2011.

[19] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? *SIGIR*, 2010.

[20] R. L. T. Santos, C. Macdonald, and I. Ounis. Aggregated search result diversification. *ICTIR*, 2011.

[21] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. *SIGIR* 2003.

[22] E. M. Voorhees. Overview of the trec 2003 question answering track. In *TREC*, 2003.

[23] X.-B. Xue, Z.-H. Zhou, and Z. M. Zhang. Improving web search using image snippets. *ACM Trans. Internet Technol.*, 8:21, 2008.

[24] J. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin.*, 76(5), 1971.

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

[26] K. Zhou, R. Cummins, M. Halvey, M. Lalmas and J. M. Jose. Assessing and Predicting Vertical Intent for Web Queries. In *ECIR*, 2012.