# A Methodology for Evaluating Aggregated Search Results

Jaime Arguello[1], Fernando Diaz[2], Jamie Callan[1], and Ben Carterette[3]

[1]Carnegie Mellon University
[2]Yahoo! Research
[3]University of Delaware

**Abstract.** Aggregated search is the task of incorporating results from different specialized search services, or *verticals*, into Web search results. While most prior work focuses on deciding *which* verticals to present, the task of deciding *where* in the Web results to embed the vertical results has received less attention. We propose a methodology for evaluating an aggregated set of results. Our method elicits a relatively small number of human judgements for a given query and then uses these to facilitate a metric-based evaluation of *any* possible presentation for the query. An extensive user study with 13 verticals confirms that, when users prefer one presentation of results over another, our metric agrees with the stated preference. By using Amazon's Mechanical Turk, we show that reliable assessments can be obtained quickly and inexpensively.

## 1 Introduction

Commercial search engines provide access to multiple specialized search services or *verticals*, such as image search, news search, local business search, and items for sale. There are two ways that users typically access vertical content. In some cases, if a user wants results from a particular vertical, they can issue the query to the vertical directly. In other cases, however, a user may not know that a vertical is relevant or may want results from multiple verticals at once. For these reasons, commercial search engines sometimes incorporate vertical results into the Web results. This is referred to as *aggregated search*.

Aggregated search can be viewed as a two-part task. Most prior work focuses on *vertical selection*—the task of predicting which verticals (if any) are relevant to a query [5, 10, 1, 6, 2]. The second task of deciding *where* in the Web results to embed the vertical results has received less attention. One possible reason for this is that a well-defined methodology for evaluating an aggregated set of results does not currently exist.

To date, aggregated results are evaluated based on user feedback, collected either implicitly (e.g., by observing clicks and skips [5, 13]) or explicitly (e.g., by directly asking users which results they prefer [15]). Existing approaches, however, focus on the integration of at most a *single* vertical into the Web results. Focusing on a single vertical simplifies evaluation by limiting the space of possible layouts or *presentations* to a manageable size. User feedback can be collected for

every possible presentation of results and, thereby, we can determine, not only whether one presentation is preferred over another, but whether one is preferred over all. This is not possible, however, if we consider many verticals (e.g., more than 10) simultaneously competing for space across the search results page. In this case, the space is too large to explore fully. The question, then, is: how can we measure the quality of *any* possible presentation for a given query? This question is central to aggregated search evaluation and is the question we address in this work.

We propose and validate a methodology for evaluating aggregated search. The goal is to elicit a relatively small number of human judgements for a given query and then to use these to evaluate *any* possible presentation of results. A central component of our methodology is the prediction of a *reference* presentation, which marks the best possible presentation that a system can produce for the given query. Given the prohibitively large space of presentations, we do not elicit human judgements on full presentations. Instead, we take a piece-wise, bottom-up approach. We collect pairwise preferences on blocks of results and use these to derive the *reference* presentation. Finally, we propose that any arbitrary presentation for the query can be evaluated based on its distance (using a rank-based metric) to the *reference*. To validate our methodology we present a user study in which we test the following hypothesis: given two alternative presentations for a query, if users prefer one over the other, then they prefer the one that is closest (in the metric space) to the *reference*.

Two resources were required to validate our methodology. First, we required a wide range of operational verticals, resembling those available to a commercial search engine. We used a set of 13 verticals developed using freely-available search APIs from various on-line services (e.g., eBay, Google, Twitter, Yahoo!, YouTube). Second, we required a pool of assessors. We used Amazon's Mechanical Turk (AMT).[1] By doing so, we show that the proposed methodology can be applied using a large pool of inexpensive, non-expert assessors and does not require an operational system with users. Therefore, it is applicable to both commercial and non-commercial environments.

## 2 Modeling Assumptions and Problem Definition

At query time, an aggregated search system issues the query to the Web search engine and to every vertical. At this point, every vertical that retrieves results is a *candidate* vertical. The task, then, is to decide *which* candidate verticals to present and *where* in the Web results to present them. The decision of where to present vertical results is subject to a set of layout constraints.

We make the following layout assumptions. First, we assume that vertical results can only be embedded in 4 positions relative to the top 10 Web results: above the first Web result, between Web results 3-4, between Web results 6-7, and below the last Web result. A similar assumption is made in prior work [13,

[1] http://www.mturk.com

15, 5]. Effectively, this divides the top 10 Web results into three blocks of results, denoted as $w_1$, $w_2$, and $w_3$. Multiple verticals can be embedded between any two Web blocks, above $w_1$, or below $w_3$. Second, we assume that users prefer to not see results from non-relevant verticals, even below $w_3$. Non-relevant verticals should be suppressed entirely. Third, we assume that if a vertical is presented, then a fixed set of its top results must be presented and must appear adjacent in the ranking. Finally, we assume that Web results are always presented and never re-ranked. That is, $w_{1-3}$ are always presented in their original order.

Given these assumptions, we can formulate the aggregation task as one of ranking blocks of Web and vertical results. A *block* is defined as a set of Web or vertical results which cannot be split in the aggregated results. If a block is presented, then all its results must be presented and must appear adjacent in the ranking. If a block is suppressed, then all its results must be suppressed. Let $\mathcal{B}_q$ denote the set of blocks associated with query $q$, which always includes all three Web blocks ($w_{1-3}$) and one block for each candidate vertical. The aggregation task is to produce a partial ranking of $\mathcal{B}_q$, denoted by $\sigma_q$. Suppressed verticals will be handled using an imaginary "end of search results" block, denoted by *eos*. Blocks that are ranked below *eos* are suppressed. We say that $\sigma_q$ is a *partial* ranking because all blocks ranked below *eos* (i.e., those that are suppressed) are effectively tied.

Our objective is an evaluation measure $\mu$ that can determine the quality of *any* possible presentation $\sigma_q$ for query $q$. The raw input to $\mu$ is a set of human judgements, denoted by $\pi_q$. Given the prohibitively large number of possible presentations, we do not elicit judgements directly on full presentations. Instead, we take a piece-wise, bottom-up approach and collect judgements on individual blocks. Prior work shows that assessor agreement is higher on document preference judgements than on absolute judgements [3]. Here, we assume this to also be true when assessing blocks of results. Therefore, we collect preference judgements on *all* pairs of blocks in $\mathcal{B}_q$. We use $\pi_q(i, j)$ to denote the number of assessors who preferred block $i$ over block $j$.

A validation of $\mu$ should be grounded on user preferences. Suppose we have two alternative presentations for a given query. If users prefer one over the other, then the preferred presentation should be the one judged superior by $\mu$.

## 3 Related Work

As previously mentioned, most prior work on aggregated search focuses on vertical selection. Li *et al.* [10] classified queries into two classes of vertical intent (*product* and *job*) and evaluated based on precision and recall for each class independently. Diaz [5] focused on predicting when to display news results (always displayed above Web results) and evaluated in terms of correctly predicted clicks and skips. Arguello *et al.* [1] focused on *single*-vertical selection, where at most a single vertical is predicted for each query. Evaluation was in terms of the number of correctly predicted queries. Diaz and Arguello [6] investigated vertical selection in the presence of user feedback. Evaluation was based on a simulated

stream of queries, where each query had at most one relevant vertical. Arguello *et al.* [2] focused on model adaptation for vertical selection and evaluated in terms of precision and recall for each vertical independently.

The above work assumes at most a *single* relevant vertical per query and, either implicitly or explicitly, assumes a fixed presentation template (e.g., *news* results are presented above Web results, if at all [5]). Users, however, may prefer the vertical results in different locations for different queries, and, in some cases, may prefer results from multiple verticals. Our proposed methodology may facilitate a more comprehensive evaluation of vertical selection. Suppose we have access to a high-quality *reference* presentation for a query. Then, for example, we might weight a false negative selection decision more if the vertical is ranked high in the *reference* and weight it less if it is ranked low.

In terms of *where* to embed the vertical results, several studies investigate user preference behavior. Sushmita *et al.* [13] investigated the effects of position and relevance on click-through behavior. They focused on presentations with one of three verticals (*images*, *news*, and *video*) slotted in one of three positions in the Web results. A positive correlation was found between both relevance and position and click-through rate. More surprisingly, perhaps, they found a bias in favor of *video* results. Users clicked more on *video* results irrespective of position and relevance. Zhu and Carterette [15] focused on user preferences with the *images* vertical and three slotting positions. They observed a strong preference towards *images* above Web results for queries likely to have image intent. From these studies, we can draw the conclusion that users care not only about *which* verticals are presented, but also *where* they are presented.

Several works elicit preference judgements on pairs of search results, as we do. Thomas and Hawking [14] validated the side-by-side comparison approach by presenting assessors with pairs of different quality (e.g., Google results 1-10/11-20, or overlapping sets 1-10/6-15). Users preferred results 1-10. Sanderson *et al.* [11] used a similar interface with Mechanical Turk to validate a set of test-collection-based metrics. NDCG agreed the most with user preferences (63% agreement overall and 82% for navigational queries).

## 4 Preference-Based Evaluation Methodology

Our objective is an evaluation measure that can determine the quality of any possible presentation $\sigma_q$ for query $q$. Our method is depicted in Fig. 1. The general idea is to evaluate presentation $\sigma_q$ based on its distance to a ground truth or *reference* presentation $\sigma_q^*$, which is generated from a set of preference judgements on pairs of Web and vertical blocks. Given query $q$, a set of blocks $\mathcal{B}_q$ is composed from Web blocks $w_{1-3}$ and from every candidate vertical. Each block-pair $i, j \in \mathcal{B}_q$ is presented to multiple assessors who are asked to state a preference. Then, we use a voting method to derive $\sigma_q^*$ from these block-wise preference judgements. Finally, we propose that any presentation $\sigma_q$ can be evaluated by using a rank-based metric to measure its distance to $\sigma_q^*$.


(a) slots (b) blocks (c) preferences (d) reference (e) evaluation

**Fig. 1.** Approach Overview.

### 4.1 Constructing the Reference Presentation

For every query $q$, we collected preference judgements on all pairs of blocks from $\mathcal{B}_q$. Each judgement consisted of a query $q$ and block pair $i, j \in \mathcal{B}_q$ presented side-by-side in random order. Assessors were given three choices: $i$ is better, $j$ is better, and both are bad. We omitted the choice that both $i$ and $j$ are equally good to prevent assessors from abstaining from difficult decisions. We interpreted the assessor selecting "both are bad" as evidence that $i$ and $j$ should be suppressed for $q$. Each triplet of the form $(q, i, j)$ was assessed by four different assessors. These preference judgements, denoted by $\pi_q$, are the raw input to the method that derives the *reference* presentation $\sigma_q^*$.

There exist many voting methods for aggregating item preference data into a single ranking. In this work, we used the Schulze voting method because of its widespread adoption and ease of implementation [12]. The general idea is the following. Let $\pi_q(i, j)$ denote the number of assessors who preferred $i$ over $j$. We say that $i$ directly defeats $j$ if $\pi(i, j) > \pi(j, i)$. That is, if more assessors preferred $i$ over $j$ than vice versa. A *beatpath* from $i$ to $j$ is a direct or indirect defeat from $i$ to $j$. An *indirect* beatpath from $i$ to $j$ is a sequence of direct defeats from $i$ to $j$. For example, if $i$ directly defeats $k$ and $k$ directly defeats $j$, then this is an *indirect* beatpath from $i$ to $j$. The strength of an indirect beatpath is the number of votes associated with its weakest direct defeat. Finally, we say that $i$ defeats $j$ if the strongest (direct or indirect) beatpath from $i$ to $j$ is stronger than the one from $j$ to $i$. Blocks are then ranked by their number of defeats.

As previously mentioned, the aggregation task is not only ranking blocks, but also deciding which vertical blocks to suppress. Suppressed verticals were handled using the imaginary *eos* block. The *eos* block was treated by the Schulze method the same as any non-imaginary block. Every time an assessor selected that both $i$ and $j$ are bad, we incremented the value of $\pi(eos, j)$ and $\pi(eos, i)$. Also, recall that we assume that Web blocks ($w_{1-3}$) are always presented and never re-ranked. This constraint was imposed by setting $\pi(eos, w_*) = \pi(w_x, w_y) = 0$, where $x > y$, and by setting $\pi(w_*, eos) = \pi(w_x, w_y) = N$, where $x < y$ and $N$ is some large number (we used $N = 1000$).

### 4.2 Measuring Distance from the Reference

Our proposed method is to evaluate *any* possible presentation $\sigma_q$ by measuring its distance to the *reference* $\sigma_q^*$. We used a rank-based distance metric. Possibly

the most widely used rank-based distance metric is Kendall's tau ($K$), which counts the number of discordant pairs between two rankings,

$$K(\sigma^*, \sigma) = \sum_{\sigma^*(i)<\sigma^*(j)} [\sigma(i) > \sigma(j)],$$

where $\sigma(i)$ denotes the rank of element $i$ in $\sigma$. Kendall's tau treats all discordant pairs equally regardless of position. In our case, however, we assume that users scan results from top-to-bottom. Therefore, we care more about a discordant pair at the top of the ranking than one at the bottom. For this reason, we used a variation of Kendall's tau proposed by Kumar and Vassilvitskii [8], referred to as *generalized* Kendall's tau ($K^*$), which can encode positional information using element weights. To account for positional information, $K^*$ models the cost of an *adjacent* swap, denoted by $\delta$. In traditional Kendall's tau, $\delta = 1$, irrespective of rank. Adjacent swaps are treated equally regardless of position. In our case, however, we would like discordant pairs at the top to be more influential. Let $\delta_r$ denote the cost of an adjacent swap between elements at rank $r-1$ and $r$. We used the DCG-like cost function proposed in Kumar and Vassilvitskii [8],

$$\delta_r = \frac{1}{\log(r)} + \frac{1}{\log(r+1)},$$

which is defined for $2 \leq r \leq n$. Given rankings $\sigma^*$ and $\sigma$, element $i$'s displacement weight $\bar{p}_i(\sigma^*, \sigma)$ is given by the average cost (in terms of adjacent swaps) it incurs in moving from rank $\sigma_q^*(i)$ to rank $\sigma_q(i)$,

$$\bar{p}_i(\sigma^*, \sigma) = \begin{cases} 1 & \text{if } \sigma^*(i) = \sigma(i) \\ \frac{p_{\sigma^*(i)} - p_{\sigma(i)}}{\sigma(i)^* - \sigma(i)} & \text{otherwise} \end{cases},$$

where $p_r = \sum_2^r \delta_r$. The $K^*$ distance is then given by,

$$K^*(\sigma^*, \sigma) = \sum_{\sigma^*(i)<\sigma^*(j)} \bar{p}_i(\sigma^*, \sigma)\bar{p}_j(\sigma^*, \sigma)[\sigma(i) > \sigma(j)].$$

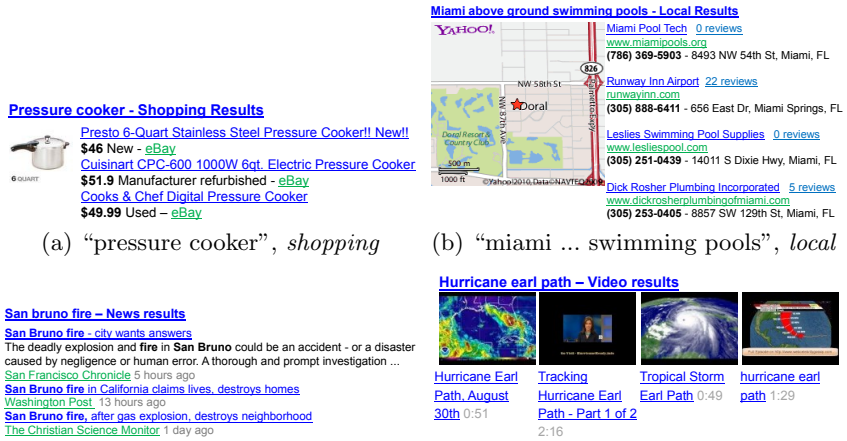
A discordant pair's contribution to the metric is equal to the product of the two element weights.

## 5 Materials and Methods

### 5.1 Verticals and Queries

We focused on a set of 13 verticals constructed using freely-available search APIs provided by eBay (*shopping*), Google (*blogs*, *books*, *weather*), Recipe Puppy (*recipes*), Yahoo! (*answers*, *finance*, *images*, *local*, *maps*, *news*), Twitter (*micro-blogs*), and YouTube (*video*). A few example vertical blocks are presented in Fig. 2. Each vertical was associated with a unique presentation of results. For

example, *news* results were associated with the article title and url, the news source title and url, and the article's publication date, and included an optional thumbnail image associated with the top result. *Shopping* results were associated with the product name and thumbnail, its condition (e.g., new, used), and its price. *Local* results were associated with the business name and url, its address and telephone number, and the number of reviews associated with it, and included a map. Each vertical was associated with a maximum number of top results (e.g., 4) from which to construct a block.




(a) "pressure cooker", *shopping* (b) "miami ... swimming pools", *local*

(c) "san bruno fire", *news* (d) "hurricane earl path", *video*

**Fig. 2.** Example vertical blocks.

Our evaluation was conducted on a set of 72 queries from two different sources: the AOL query log and Google Trends. Google Trend queries cover recent events and topics currently discussed in news articles, blogs, and on Twitter (e.g., "us open fight"). AOL queries cover more persistent topics likely to be relevant to verticals such as *local* (e.g., "cheap hotels in anaheim ca"), *recipe* (e.g., "cooking ribs"), and *weather* (e.g., "marbella weather"). Queries were selected manually in order to ensure coverage for our set of 13 verticals.

### 5.2 Preference Judgements on Block-Pairs

While collecting block-pair judgements, in addition to the query, assessors were given a topic description to help disambiguate the user's intent. In a preliminary experiment, we observed an improvement in inter-annotator agreement from giving assessors topic descriptions. We were careful, however, to not explicitly mention vertical intent. For example, for the query "pressure cooker", we stated: "The user plans to buy a pressure cooker and is looking for product information." We did not say: "The user is looking for shopping results." The assessments were

conducted using Amazon's Mechanical Turk (AMT). Turkers were compensated 0.01 US$ for each judgement.

Following Sanderson *et al.* [11], quality control was done by including 150 "trap" HITs (a Human Intelligence Task is a task associated with AMT). Each trap HIT consisted of a triplet $(q, i, j)$ where either $i$ 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 judgement. Assessors who failed more than a couple of trap HITs were removed from the judgement pool.

## 6 Assessor Agreement on Block-Pair Judgements

Of the 120 assessors who contributed HITs, 2 had their assessments removed from the assessment pool due to failing more than 2 trap HITs. For the remaining 118/120, participation followed a power law distribution—about 20% (24/118) of the assessors completed about 80% (9,856/12,293) of our HITs.

We report inter-annotator agreement in terms of Fleiss' Kappa ($\kappa_{\mathrm{f}}$) [7] and Cohen's Kappa ($\kappa_{\mathrm{c}}$) [4], both which correct for agreement due to chance. Fleiss' Kappa measures the (chance-corrected) agreement between *any* pair of assessors over a set of triplets. Cohen's Kappa measures the (chance-corrected) agreement between a *specific* pair of assessors over a *common* set of triplets. For our purpose, Fleiss' Kappa is convenient because it ignores the identity of the assessor-pair. It is designed to measure agreement over instances labeled by different (even disjoint) sets of assessors. However, precisely because it ignores the identity of the assessor-pair, it is dominated by the agreement between the most active assessors, which we know to be a selected few. To compensate for this, in addition to Fleiss' Kappa, we present the Cohen's Kappa agreement for all pairs of assessors who labeled at least 100 triplets in common.

The Fleiss' Kappa agreement over *all* triplets was $\kappa_{\mathrm{f}} = 0.656$, which is considered *substantial* agreement based on Landis and Koch [9]. In terms of Cohen's Kappa agreement, there were 25 pairs of assessors with at least 100 triplets in common. Of these, 5 (20%) had *moderate* agreement ($0.40 < \kappa_c \leq 0.60$), 16 (64%) had *substantial* agreement ($0.60 < \kappa_c \leq 0.80$), and the remaining 4 (16%) had *perfect* agreement ($0.80 < \kappa_c \leq 1.00$). Overall, assessor agreement on block-pairs was high. We view this as evidence that assessors did not have difficulty providing preferences for pairs of Web and vertical blocks.

## 7 Empirical Analysis and Validation

A desirable property of any evaluation measure is that it should correlate with user preference. We conducted a user study to test whether our metric (the $K^*$ distance between $\sigma_q$ and $\sigma_q^*$) satisfies this criterion. Users were shown pairs of presentations side-by-side (along with the query and its description) and were asked to state a preference ("left is better", "right is better"). We assumed that assessors would have difficulty deciding between two bad presentations. Therefore, to reduce cognitive load, we also included a "both are bad" option. Our

hypothesis is that our metric will agree with the stated preference. Significance was tested using a sign test, where the null hypothesis is that the metric selects the preferred presentation randomly with equal probability.

Conducting this analysis requires a method for selecting pairs of presentations to show assessors. One alternative is to sample pairs uniformly from the set of all presentations. However, we were particularly interested in pairs of presentations from *specific* regions of the metric space. For example, is the metric correlated with user preference when one presentation is presumably high-quality (close to the reference) and the other is low-quality (far from the reference). Is it correlated when *both* presentations are presumably high-quality or when *both* are low-quality? To investigate these questions, we sampled presentation-pairs using a binning approach. For each query, presentations were divided into three bins: a high-quality bin ($\mathcal{H}$), a medium-quality bin ($\mathcal{M}$), and a low-quality bin ($\mathcal{L}$). The binning was done based on the metric value. The metric distribution is such that this produces bins where $|\mathcal{H}| < |\mathcal{M}| < |\mathcal{L}|$. The $\mathcal{H}$ bin is the smallest and contains those presentations that are nearest to $\sigma_q^*$. The $\mathcal{L}$ bin is the largest and contains those presentations that are furthest from $\sigma_q^*$. For each query, we sampled 4 presentation-pairs from each bin-combination ($\mathcal{H}$-$\mathcal{H}$, $\mathcal{H}$-$\mathcal{M}$, $\mathcal{H}$-$\mathcal{L}$, $\mathcal{M}$-$\mathcal{M}$, $\mathcal{M}$-$\mathcal{L}$, and $\mathcal{L}$-$\mathcal{L}$) and collected 4 judgements per presentation-pair. This resulted in 1,728 presentation-pairs and 6,912 judgements. For this analysis, we also used Amazon's Mechanical Turk.

### 7.1 Results

Assessor agreement on presentation-pairs was $\kappa_{\mathrm{f}} = 0.216$, which is considered *fair* agreement [9]. Of all 1,728 presentation-pairs, only 1,151 (67%) had a majority preference of *at least* 3/4 and only 462 (27%) had a perfect 4/4 majority preference. It is perhaps not surprising that assessor agreement was lower on presentation-pairs than on block-pairs. Agreement on presentation-pairs requires that assessors make similar assumptions about the cost of different types of errors: a false-positive (displaying a non-relevant vertical), a false-negative (suppressing a relevant vertical), and a ranking error (displaying a relevant vertical in the wrong position). Assessors may require more instruction in order to improve agreement on presentation-pairs. Alternatively, more than 4 assessors may be required to see greater convergence.

Given this low level of inter-assessor agreement, rather than focus on the metric's agreement with each individual preference, we focus on its agreement with the *majority* preference. We present results for two levels of majority preference: a majority preference of 3/4 or greater and a perfect (4/4) majority preference. These results are presented in Table 1. The "pairs" column shows the number of presentation pairs for which the level of majority preference was observed. The "% agreement" column shows the percentage of these pairs for which the metric agreed with the majority preference.

The metric's agreement with the majority preference was 67% on pairs where at least 3/4 assessors preferred the same presentation and 73% on pairs where all (4/4) assessors preferred the same presentation (both significant at the $p < 0.005$

**Table 1.** Metric agreement with majority preference. Significance is denoted by † and ‡ at the $p < 0.05$ and $p < 0.005$ level, respectively

| bins | majority preference | pairs | % agreement |
|---|---|---|---|
| all | 3/4 preference | 1151 | 67.07%‡ |
| $\mathcal{H}$-$\mathcal{H}$ | 3/4 or greater | 164 | 60.37%‡ |
| $\mathcal{H}$-$\mathcal{M}$ | 3/4 or greater | 210 | 81.90%‡ |
| $\mathcal{H}$-$\mathcal{L}$ | 3/4 or greater | 204 | 84.31%‡ |
| $\mathcal{M}$-$\mathcal{M}$ | 3/4 or greater | 184 | 57.61%† |
| $\mathcal{M}$-$\mathcal{L}$ | 3/4 or greater | 187 | 50.80% |
| $\mathcal{L}$-$\mathcal{L}$ | 3/4 or greater | 202 | 63.37%‡ |
| all | 4/4 | 462 | 72.51%‡ |
| $\mathcal{H}$-$\mathcal{H}$ | 4/4 | 47 | 65.96%† |
| $\mathcal{H}$-$\mathcal{M}$ | 4/4 | 95 | 87.37%‡ |
| $\mathcal{H}$-$\mathcal{L}$ | 4/4 | 97 | 91.75%‡ |
| $\mathcal{M}$-$\mathcal{M}$ | 4/4 | 75 | 58.67% |
| $\mathcal{M}$-$\mathcal{L}$ | 4/4 | 71 | 54.93% |
| $\mathcal{L}$-$\mathcal{L}$ | 4/4 | 77 | 63.64%† |

level). Agreement with each individual preference (not in Table 1) was 60% (also significant at the $p < 0.005$ level).

One important trend worth noting is that the metric's agreement with the majority preference was higher on pairs where there was greater consensus between assessors. Overall, the metric's agreement with the majority preference was higher on presentation-pairs that had a perfect (4/4) majority preference than on pairs that had a (3/4) majority preference or greater. This is a positive result if we primarily care about pairs in which one presentation was strongly preferred over the other.

A similar trend was also observed across bin-combinations. The metric's agreement with the majority was the highest on $\mathcal{H}$-$\mathcal{M}$ and $\mathcal{H}$-$\mathcal{L}$ pairs (82-92%). These were also the bin-combinations with the highest inter-assessor agreement ($\kappa_{\text{f}} = 0.290$ for $\mathcal{H}$-$\mathcal{M}$ and $\kappa_{\text{f}} = 0.303$ for $\mathcal{H}$-$\mathcal{L}$).[2] This means that, on average, $\sigma_q^*$ was good. Assessors strongly preferred presentations close to $\sigma_q^*$ over presentations far from $\sigma_q^*$ in terms of $K^*$.

The metric was less predictive for $\mathcal{H}$-$\mathcal{H}$ pairs (60-66%). However, inter-assessor agreement on these pairs was also low ($\kappa_{\text{f}} = 0.066$, which is almost random). It turns out that most $\mathcal{H}$-$\mathcal{H}$ pairs had identical top-ranked blocks. This is because the $\mathcal{H}$ bin corresponds to those presentations closest to $\sigma_q^*$ based on $K^*$, which focuses on discordant pairs in the top ranks. About half of all $\mathcal{H}$-$\mathcal{H}$ pairs had the same top 3 blocks and *all* pairs had the same top 2 blocks. The low inter-assessor agreement may be explained by users primarily focusing on the top results, perhaps rarely scrolling down to see the results below the "fold". Alternatively, it may be that assessors have a hard time distinguishing

[2] Inter-assessor agreement across bin-combinations is also reflected in column "pairs".

between good presentations. Further experiments are required to determine the exact cause of disagreement. The metric was also less predictive for $\mathcal{M}$-$\mathcal{M}$, $\mathcal{M}$-$\mathcal{L}$, and $\mathcal{L}$-$\mathcal{L}$ pairs. Again, inter-assessor agreement was also lower for pairs in these bin-combinations ($\kappa_{\mathrm{f}} = 0.216$, $\kappa_{\mathrm{f}} = 0.179$, and $\kappa_{\mathrm{f}} = 0.237$, respectively). Inter-assessor agreement (and the metric's agreement with the majority preference) was lower when neither presentation was of high quality (close to $\sigma_q^*$).

We examined the queries for which the metric's agreement with the majority preference was the lowest. In some cases, assessors favored presentations with a particular vertical ranked high, but the vertical was not favored in the block-pair judgements (therefore, it was ranked low or suppressed in $\sigma_q^*$). For example, for the query "ihop nutritional facts", assessors favored presentations with *images* ranked high. For the query "nikon coolpix", assessors favored presentations with *shopping* ranked high. For the queries "san bruno fire", "learn to play the banjo", "miss universe 2010", and "us open fight", assessors favored presentations with *video* ranked high. All three verticals (*images*, *shopping*, and *video*) are visually appealing (i.e., their blocks include at least one image). Prior research found a click-through bias in favor of visually appealing verticals (e.g., *video*) [13]. It may be that this type of bias affected assessors more on presentation-pairs (i.e., where the vertical is embedded within less visually appealing results) than on block-pairs (where the vertical is shown in isolation). If accounting for such a bias is desired, then future work might consider incorporating more context into the block-pair assessment interface. One possibility could be to show each block embedded in the same position within the same set of results.

## 8 Conclusion

We described a new methodology for evaluating aggregated search results. The idea is to use preference judgements on block-pairs to derive a *reference* presentation for the query and then to evaluate alternative presentations based on their distance to the *reference*. The approach has several advantages. First, with only a relatively small number of assessments per query, we can evaluate *any* possible presentation of results. This is not only useful for evaluation, but may be useful for learning and optimization. Second, the approach is general. We used a particular interface for assessing block-pairs, a particular voting method for deriving the *reference*, and a particular rank-similarity metric for measuring distance from the *reference*. Future work may consider others. Third, we showed that reliable block-pair assessments can be collected from a pool of inexpensive assessors. Finally, we presented a user study to empirically validate our metric. Assessors were shown pairs of presentations and asked to state a preference. Overall, the metric's agreement with the majority preference was in the 67-73% range. Agreement was in the 82-92% range on those pairs where there was greater consensus between assessors.

In terms of future work, some open questions remain. Assessor agreement on presentation-pairs was low. Further experiments are needed to understand why. It may be, for example, that users assign a different value to different

types of errors (false positives, false negatives, ranking errors). Also, in some cases, assessors favored a particular vertical only when seen within the context of *other* results. There may be preferential biases that affect presentation-pair judgements more than block-pair judgements.

## 9 Acknowledgments

This work was supported in part by the NSF grants IIS-0916553, IIS-0841275, and IIS-1017026 as well as a generous gift from Yahoo! through its Key Scientific Challenges program. Any opinions, findings, conclusions, and recommendations expressed in this paper are the authors' and do not necessarily reflect those of the sponsors.

## References

1. J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. In *SIGIR 2009*, pages 315–322. ACM, 2009.
2. J. Arguello, F. Diaz, and J.-F. Paiement. Vertical selection in the presence of unlabeled verticals. In *SIGIR 2010*, pages 691–698. ACM, 2010.
3. B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: preference judgments for relevance. In *ECIR 2008*, pages 16–27, 2008.
4. J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.
5. F. Diaz. Integration of news content into web results. In *WSDM 2009*, pages 182–191. ACM, 2009.
6. F. Diaz and J. Arguello. Adaptation of offline vertical selection predictions in the presence of user feedback. In *SIGIR 2009*, pages 323–330. ACM, 2009.
7. J. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.
8. R. Kumar and S. Vassilvitskii. Generalized distances between rankings. In *WWW 2010*, pages 571–580. ACM, 2010.
9. J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.
10. X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *SIGIR 2008*, pages 339–346. ACM, 2008.
11. M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *SIGIR 2010*, pages 555–562. ACM, 2010.
12. M. Schulze. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Social Choice and Welfare*, July 2010.
13. S. Sushmita, H. Joho, M. Lalmas, and R. Villa. Factors affecting click-through behavior in aggregated search interfaces. In *CIKM 2010*, pages 519–528. ACM, 2010.
14. P. Thomas and D. Hawking. Evaluation by comparing result sets in context. In *CIKM 2006*, pages 94–101. ACM, 2006.
15. D. Zhu and B. Carterette. An analysis of assessor behavior in crowdsourced preference judgements. In *SIGIR Workshop on Crowdsourcing for Search Evaluation*, pages 21–26. ACM, 2010.