Evaluating Intuitiveness of Vertical-Aware Click Models

Aleksandr Chuklin[*]
University of Amsterdam
Amsterdam, The Netherlands
a.chuklin@uva.nl

Ke Zhou[†]
University of Edinburgh
Edinburgh, UK
ke.zhou@ed.ac.uk

Anne Schuth
University of Amsterdam
Amsterdam, The Netherlands
anne.schuth@uva.nl

Floor Sietsma
University of Amsterdam
Amsterdam, The Netherlands
fsietsma@uva.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
derijke@uva.nl

ABSTRACT

Modeling user behavior on a search engine result page is important for understanding the users and supporting simulation experiments. As result pages become more complex, click models evolve as well in order to capture additional aspects of user behavior in response to new forms of result presentation.

We propose a method for evaluating the *intuitiveness* of vertical-aware click models, namely the ability of a click model to capture key aspects of aggregated result pages, such as vertical selection, item selection, result presentation and vertical diversity. This method allows us to isolate model components and therefore gives a multi-faceted view on a model's performance. We argue that our method can be used in conjunction with traditional click model evaluation metrics such as log-likelihood or perplexity. In order to demonstrate the power of our method in situations where result pages can contain more than one type of vertical (e.g., *Image* and *News*) we extend the previously studied Federated Click Model such that it models user clicks on such pages. Our evaluation method yields non-trivial yet interpretable conclusions about the intuitiveness of click models, highlighting their strengths and weaknesses.

Categories and Subject Descriptors

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

Keywords

Aggregated search; federated search; click models; evaluation

1. INTRODUCTION

The problem of predicting user clicks has recently gained considerable interest. A *click model* is a probabilistic model of user behavior on a search engine result page. It is used to facilitate simulated experiments when real click data is limited or simply unavailable (see, e.g., [4]). In addition, the parameters of click models inferred from real clicks help us understand user behavior [6] and estimate the relevance of documents shown to the user [1].

Several *vertical-aware* click models have recently been developed (e.g., [3, 5]). These models aim to capture different aspects of user interaction with a so-called *aggregated* or *vertical* search system. The result page of an aggregated search (AS) system usually contains, in addition to regular *Web* results, heterogeneous results federated from different sub-collections or verticals (e.g., *Image*, *Video* or *News*), which are then presented in a grouped fashion. These vertical results influence user behavior in new ways [3].

Traditionally, click models have been evaluated on their ability to predict future clicks from past observations [3, 6]. More is required for the evaluation of vertical-aware click models, as we need to assess their ability to capture the peculiarities of the user behavior on the AS result page such as *attention bias*. We adapt the intuitiveness test proposed by Sakai [9] to evaluate vertical-aware click models. While Sakai's idea has been adopted before, namely for the setting of aggregated search metrics [14], our contribution is new in that we apply it to evaluate click models instead of metrics.

The main research question that we aim to answer is: *How can we evaluate the ability of a click model to capture key aspects of a vertical (aggregated) search system?* In the process of answering this question we learn that none of the existing click models are designed to deal with a result page containing multiple verticals (e.g., both *News* and *Video*). We introduce such a model and evaluate it using our evaluation method.

2. METHOD

Sakai [9] proposes a way of quantifying "which metric is more intuitive." This method has been applied to understanding aggregated search metrics in [14], where four key factors of aggregated search systems are listed: vertical selection (VS), item selection (IS), result presentation (RP) and vertical diversity (VD). The authors measure the preference agreement of a given aggregated search metric with a "basic" single-component metric for each factor; they also assess the ability of a metric to capture the combination of these factors.

The main contribution of our work is that we adapt the intuitiveness test to evaluate vertical-aware *click models* instead of aggregated search metrics. In order to apply the intuitiveness test to click models, we use a simulation setup and proceed as follows. We run a click model *CM* to simulate user clicks[1] and report the total number of clicks (CTR) produced by the simulated user as a metric score for a given ranking. We then compare AS systems by the number

[*] Now at Google Switzerland.

[†] Part of this work was conducted while at University of Glasgow.

[1] The code we use to simulate clicks is available as a part of Lerot [10] at `https://bitbucket.org/ilps/lerot`.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
ACM 978-1-4503-2257-7/14/07.
http://dx.doi.org/10.1145/2600428.2609513

```
$Disagree = 0$; $Correct_1 = 0$; $Correct_2 = 0$;
foreach pair of runs $(r_1, r_2)$ do
    foreach TREC topic $t$ do
        $\delta_1 = CTR_{CM_1}(t, r_1) - CTR_{CM_1}(t, r_2)$;
        $\delta_2 = CTR_{CM_2}(t, r_1) - CTR_{CM_2}(t, r_2)$;
        $\delta_{GS} = M_{GS}(t, r_1) - M_{GS}(t, r_2)$;
        if $(\delta_1 \times \delta_2) < 0$ then // $CM_1$ and $CM_2$ disagree
            $Disagree$++;
            if $\delta_1 \times \delta_{GS} \geq 0$ then // $CM_1$ and $M_{GS}$ agree
                $Correct_1$++;
            if $\delta_2 \times \delta_{GS} \geq 0$ then // $CM_2$ and $M_{GS}$ agree
                $Correct_2$++;
$Intuitiveness(CM_1|CM_2, M_{GS}) = Correct_1/Disagree$;
$Intuitiveness(CM_2|CM_1, M_{GS}) = Correct_2/Disagree$;
```

Algorithm 1: Computing the intuitiveness scores of click models CM_1 and CM_2 based on preference agreement with a gold standard metric M_{GS}.

of clicks they receive according to a click model CM, like in A/B-testing experiments.[2] The outcome of this AS system comparison determines the intuitiveness of the underlying click model.

Algorithm 1 shows our intuitiveness test algorithm. The algorithm computes relative intuitiveness scores for a pair of click models CM_1 and CM_2 and a gold standard metric M_{GS}. The latter represents a basic property that a candidate metric should satisfy. We consider not one but four metrics as our gold standards, one for each aggregated search factor; the same metrics were used by [14]. These gold standards are intentionally kept simple. They should be agnostic to differences across metrics (e.g., different position-based discounts); their purpose is to separate out and test single factor properties of more complex click models. The four gold standard metrics are: (a) VS: vertical precision; (b) VD: vertical recall; (c) IS: mean precision of vertical result items; and (d) RP: Spearman's rank correlation with a "perfect" AS reference page.

We first obtain all pairs of AS result pages for which CM_1 and CM_2 disagree about which result page should get more clicks. Out of these disagreements, we count how often each click model's CTR scores agree with the gold standard metric(s). The click model that concords more with the gold standard metric(s) is considered to be more "intuitive." An ideal click model should be consistent with all four gold standards; we therefore add an additional step to Algorithm 1 by counting how often the model agrees with a subset or all the four gold standards at the same time.

When compared to traditional perplexity-based click model evaluation [3, 6], our method has the following advantages: (1) it allows for assessments of *individual* model components, separating their contribution to the model's performance; (2) it assigns *explanatory* scores that allow us to assess the ideas underlying a click model; and (3) it allows us to make use of public test collections and obtain *re-usable* scores without need to access a user click log.

3. EXPERIMENTAL SETUP

In this section we present the click models (both traditional and vertical-aware) that we use to demonstrate our intuitiveness evaluation method. We first introduce the models and then specify their parameters as well as the aggregated search dataset that we use.

3.1 Click Models

Traditional Click Models. We start with the simple traditional click models and then move to the more complex vertical-aware models. In order to model the user's behavior on a search engine result page (SERP), a click model usually employs two sets of binary random variables: E_i (examination), which equals 1 if the user examines the i-th document snippet, C_i (click), which equals 1 if the user clicks the i-th document link. The Random Click Model (RCM) assumes that a document is clicked with probability $p = 0.5$ regardless of document position and its perceived relevance: $P(C_i = 1) = p$. If we take into account the fact that the documents lower in the SERP have a lower chance of being examined [12], we obtain a Position-Based Model (PBM): $P(C_i = 1) = P(C_i = 1 \mid E_i = 1) \cdot P(E_i = 1)$, where the examination probability $P(E_i = 1) = p^{i-1}$ and the probability of a click given examination is approximated using relevance labels:[3] $P(C_i = 1 \mid E_i = 1) = r_i$; a similar model was used as a baseline model in [1]. Following Zhang et al. [12], we set $p = 0.73$.

Vertical-Aware Click Models. Chen et al. [3] found that about 15% of the search result pages contain more than one type of vertical. Since this is a significant fraction of the search traffic, we want to adequately evaluate click models that capture user behavior in such multi-vertical settings. In order to demonstrate how our method rates such models, we introduce a multi-vertical Federated Click Model ($mFCM$), a generalization of the Federated Click Model (FCM) by Chen et al. [3, 4] in which we allow different vertical types, each with its own influence on examination probabilities.

As in [3], $P(E_i = 1)$ is influenced by the distance to different verticals on the page and the *attention bias* caused by these verticals. If there is no attention bias present, the examination probability ϕ_i depends only on the rank of the document i:

$$P(E_i = 1 \mid A) = \phi_i + (1 - \phi_i)\beta_i(A) \tag{1}$$

$$P(C_i = 1 \mid E_i = 0) = 0 \tag{2}$$

$$P(C_i = 1 \mid E_i = 1) = r_i. \tag{3}$$

Here, A is the vector of independent binary random variables A_j, attention bias values for each vertical $vert_j$. The influence of vertical documents on the examination probability of a document i is represented by a function $\beta_i(A)$. We set it to 1 if document i is the vertical document itself and decrease it as document i is further away from the vertical documents [3]. According to Chen et al. [3], the decrease should depend on the vertical type j, so we introduce parameters γ_j that depends solely on the vertical type j:

$$\beta_i(A) = \min\left(1, \max_{\{j : A_j = 1\}} \frac{1}{|dist_j(i)| + \gamma_j}\right), \tag{4}$$

where $dist_j(i)$ is the distance from document i to the nearest document that belongs to $vert_j$ [3, 4].

If we do not distinguish between different verticals in (4), i.e. set $\gamma_j = \gamma$ for all j, and also assume that for a vertical j, its attention bias A_j is determined only by its position on the page, i.e., $P(A_j = 1) = hpos_{vert_j}$, we obtain the FCM model exactly as it was used in [4]. If we do assume that γ_j takes different values for different j, we get the model that we call here $mFCM\text{-}NO$.

In order to further distinguish different verticals we use the *vertical orientation* of the user, the probability that users prefer a certain vertical to general web results [13, 14]. We write $orient(vert_j, q)$ to denote the orientation of the user towards the type of $vert_j$, given query q. Having orientation values, we can further improve our click model by refining the estimation of attention bias:

$$P(A_j = 1) = orient(vert_j, q) \cdot hpos_{vert_j}. \tag{5}$$

[2] Alternatively, one can perform an interleaving comparison [2], but this is beyond the scope of the paper.

[3] For simplicity we use binary relevance labels, following [4].

The model defined by equations (1)–(5) is called $mFCM$. The simpler $mFCM\text{-}NO$ model that does not use vertical orientation ("NO" for "no orientation"), is also of interest, since vertical orientation values are not always available and it is important to understand their contribution.

3.2 Data and Parameter Settings

For the $mFCM$ model we instantiate the γ, ϕ and $hpos$ parameters similar to [4]. We set γ to 0.1 for multimedia verticals such as *News* or *Blogs* and 0.2 for text-based verticals such as *Image* or *Video* to resemble click heatmaps reported by [3] for the corresponding vertical types. We also set $hpos$ = [.95, .9, .85, .8, .75, .7, .3, .25, .2, .15] for multimedia verticals as in [4] (assuming the user cannot see documents below rank 6 without scrolling), and for text verticals $hpos = [.95, .3, .25, .15, .10, .05, .05, .05, .05, .05]$, since Chen et al. [3] suggest that a text vertical, unlike multimedia verticals, does not substantially influence user clicks if it is not at the top of the page; this is also supported by [11]. As in [4], ϕ equals $[.68, .61, .48, .34, .28, .2, .11, .1, .08, .06]$ based on the eye fixation probabilities reported by Joachims et al. [7].

To complete the experimental setup we need to specify the aggregated search systems and document dataset that we use. We use simulated aggregated systems from [14], which are built by systematically varying the quality of key aggregated search components. Specifically, we use 4 state-of-the-art VS systems, 3 ranking functions for selecting vertical items for IS and 3 ways to embed vertical result blocks on the final AS pages (RP). In total, we have simulated 36 AS systems ($4 \times 3 \times 3$). As a document collection we use a public aggregated search dataset [8] for which relevance judgements of documents and vertical orientation preference judgments are available for each topic. There are 50 test topics in our collection, so with 36 simulated AS systems runs we have a total of $C_{36}^2 = 630$ run pairs and $50 \cdot 630 = 31,500$ pairs of result pages.

4. RESULTS

We report on the intuitiveness scores computed for a variety of click models, using Algorithm 1. For each click model we test intuitiveness with respect to the four AS factors individually, as well as the ability to capture a combination of multiple AS factors. The models that we test are: $mFCM$, $mFCM\text{-}NO$, FCM, PBM, RCM, all of which are described in Section 3.1. Table 1 lists our results. For every gold standard metric and every pair of click models, we give the intuitiveness scores of both models and the percentage of result page pairs for which the models disagree.

For example, Table 1 (a) shows that if we compare $mFCM$ and $mFCM\text{-}NO$ in terms of the component VS (the ability to select relevant verticals), there are 14.7% ($4,620$ out of $31,500$ pairs) disagreements. The intuitiveness score for $mFCM$ is 0.870, which is the fraction of these disagreements for which $mFCM$ agrees with the gold standard metric. The score for $mFCM\text{-}NO$ is only 0.833, so $mFCM$ is more likely to agree with VS metric than $mFCM\text{-}NO$. Note that the scores of two competing models do not add up to 1; when the gold standard judges two result pages to be equally good, both click models agree with the gold standard. That is also why the scores for a very simple RCM are relatively high in Table 1 (a). We can also observe that as two click models differ more, the percentage of disagreements increases. For instance, the more complex click models tend to have a substantial disagreement with the random click model RCM.

Let "$CM_1 > CM_2$" denote the relationship "click model CM_1 statistically significantly outperforms click model CM_2 in terms of concordance with a given gold-standard metric." Our findings can be summarized as follows:

- VS: $mFCM > mFCM\text{-}NO, FCM$;
- VD: $mFCM, mFCM\text{-}NO > PBM > FCM > RCM$;
- IS: $mFCM > mFCM\text{-}NO > FCM > PBM > RCM$;
- RP: $FCM > PBM > mFCM, mFCM\text{-}NO > RCM$;
- VS and IS: $mFCM > mFCM\text{-}NO > FCM > PBM > RCM$;
- VS, IS, VD: $mFCM > mFCM\text{-}NO > FCM > PBM > RCM$;
- VS, IS, RP, VD (all four metrics): $mFCM > PBM > RCM$, $FCM > mFCM\text{-}NO > PBM > RCM$.

For single-component evaluation, $mFCM$ outperforms the other models on VS, VD and IS, with $mFCM\text{-}NO$ as a second-best alternative. The same holds for the combinations VS + IS and VS + IS + VD. For RP, FCM performs best, with PBM ranking second. The RP factor is measured as correlation with a "perfect" result page in which highly oriented verticals are put on top. However, this order does not necessarily emit the maximum number of clicks in FCM-like click models. For example, if there is a vertical lower on the page that attracts a lot of attention, it may be better to place the relevant document just above or below this vertical. The table suggests that the intrinsic "optimal" result orders for $mFCM$ and $mFCM\text{-}NO$ are further from the "perfect" order than PBM's. When we look at all gold standard metrics combined, FCM and $mFCM$ are almost equally good, with $mFCM\text{-}NO$ again as a second-best alternative.

Our evaluation method implies that $mFCM$ captures the VS, IS and VD factors very well, better than any other model we tested. Even without orientation values, $mFCM\text{-}NO$ is able to capture these factors. $mFCM$ performs worse at capturing result presentation as measured by our RP metric, which is unsurprising as the multiple vertical click model focuses less on putting relevant results on top and better accounts for attention bias caused by multiple vertical blocks. This shows that our intuitiveness evaluation method is able to draw non-trivial detailed conclusions about model's performance. These conclusions do not contradict our prior knowledge about the click models and can always be explained.

5. CONCLUSION AND DISCUSSION

We introduced an evaluation method that can be used to assess a vertical-aware click model's ability to capture key components of an aggregated search system and demonstrated it using different vertical-aware as well as traditional click models. We also showed that click models that account for multiple vertical blocks within a single result page typically get higher intuitiveness scores, which indicates that our evaluation method measures the right thing.

One limitation of the work presented here is that we do not use raw click data to infer the parameters of the click models we experiment with. However, we set these parameters using previous work that does use real click and eye gaze data [3, 7, 11].

As a direction for future work we want to compare the findings of the intuitiveness test with conventional model performance tests (e.g., perplexity of click prediction) and see whether good intuitiveness scores also imply good click prediction results and vice-versa.

Acknowledgements. We thank Maria-Hendrike Peetz, Süleyman Yürekli and our anonymous reviewers for their feedback on the paper. This research was partially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nrs 288024 and 312827, the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 612.066.930, 640.006.013, the Center for Creation, Content and Technology (CCCT), the QuaMerdes project funded by the CLARIN-nl program, the TROVe project funded by the CLARIAH program, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.105, the Yahoo! Faculty Research and Engagement Program, the Microsoft Research PhD program, and the HPC Fund.

Table 1: **Intuitiveness test results. For each pair of click models, the higher score is shown in bold, with the fraction of disagreements in parentheses. Results (a)–(d) show click model performance w.r.t. individual AS components; results (e)–(g) concern a click model's ability to capture multiple components. Significant differences (sign test) are indicated with △ ($\alpha = 0.05$) and ▲ ($\alpha = 0.01$).**

Evaluation Criteria		$mFCM\text{-}NO$	FCM	PBM	RCM
(a). **(VS)** gold standard: vertical selection precision	$mFCM$	**0.870**/0.833△ (14.7%)	**0.865**/0.842△ (18.2%)	**0.873**/0.856 (21.5%)	0.851/**0.865** (44.5%)
	$mFCM\text{-}NO$	-	0.843/**0.850** (17.0%)	0.868/**0.877** (23.1%)	0.845/**0.870** (44.3%)
	FCM	-	-	0.879/**0.885** (22.0%)	0.849/**0.871** (44.2%)
	PBM	-	-	-	0.848/**0.868** (45.3%)
(b). **(VD)** gold standard: vertical recall	$mFCM$	**0.819**/0.812 (14.7%)	**0.879**/0.713▲ (18.2%)	**0.832**/0.748▲ (21.5%)	**0.860**/0.678▲ (44.5%)
	$mFCM\text{-}NO$	-	**0.884**/0.715▲ (17.0%)	**0.817**/0.743▲ (23.1%)	**0.858**/0.677▲ (44.3%)
	FCM	-	-	0.763/**0.819**▽ (22.0%)	**0.822**/0.708▲ (44.2%)
	PBM	-	-	-	**0.828**/0.688▲ (45.3%)
(c). **(IS)** gold standard: mean precision of vertical retrieved items	$mFCM$	**0.765**/0.732△ (14.7%)	**0.754**/0.691▲ (18.2%)	**0.832**/0.515▲ (21.5%)	**0.918**/0.349▲ (44.5%)
	$mFCM\text{-}NO$	-	**0.745**/0.706△ (17.0%)	**0.815**/0.542▲ (23.1%)	**0.912**/0.352▲ (44.3%)
	FCM	-	-	**0.805**/0.549▲ (22.0%)	**0.902**/0.357▲ (44.2%)
	PBM	-	-	-	**0.828**/0.423▲ (45.3%)
(d). **(RP)** gold standard: Spearman Correlation with "perfect" aggregated search page	$mFCM$	**0.601**/0.592 (14.7%)	0.493/**0.691**▼ (18.2%)	0.575/**0.649**▼ (21.5%)	**0.643**/0.551▲ (44.5%)
	$mFCM\text{-}NO$	-	0.477/**0.702**▼ (17.0%)	0.569/**0.644**▼ (23.1%)	**0.642**/0.553▲ (44.3%)
	FCM	-	-	**0.653**/0.576▲ (22.0%)	**0.683**/0.513▲ (44.2%)
	PBM	-	-	-	**0.650**/0.527▲ (45.3%)
(e). **(VS + IS)** gold standard: vertical selection precision AND vertical item mean precision	$mFCM$	**0.666**/0.606△ (14.7%)	**0.651**/0.584▲ (18.2%)	**0.728**/0.432▲ (21.5%)	**0.782**/0.294▲ (44.5%)
	$mFCM\text{-}NO$	-	**0.630**/0.608△ (17.0%)	**0.709**/0.474▲ (23.1%)	**0.772**/0.301▲ (44.3%)
	FCM	-	-	**0.714**/0.486▲ (22.0%)	**0.766**/0.303▲ (44.2%)
	PBM	-	-	-	**0.700**/0.362▲ (45.3%)
(f). **(VS + IS + VD)** gold standard: vertical selection precision AND vertical item mean precision AND vertical recall	$mFCM$	**0.567**/0.522△ (14.7%)	**0.585**/0.456▲ (18.2%)	**0.605**/0.347▲ (21.5%)	**0.680**/0.248▲ (44.5%)
	$mFCM\text{-}NO$	-	**0.569**/0.469▲ (17.0%)	**0.580**/0.372▲ (23.1%)	**0.673**/0.253▲ (44.3%)
	FCM	-	-	**0.568**/0.430▲ (22.0%)	**0.649**/0.269▲ (44.2%)
	PBM	-	-	-	**0.601**/0.300▲ (45.3%)
(g). **(VS + IS + RP + VD)** gold standard: ALL single-component metrics	$mFCM$	0.372/**0.373** (14.7%)	0.346/**0.366** (18.2%)	**0.394**/0.257▲ (21.5%)	**0.485**/0.164▲ (44.5%)
	$mFCM\text{-}NO$	-	0.350/**0.370**▽ (17.0%)	**0.390**/0.263▲ (23.1%)	**0.488**/0.164▲ (44.3%)
	FCM	-	-	**0.414**/0.269▲ (22.0%)	**0.486**/0.155▲ (44.2%)
	PBM	-	-	-	**0.435**/0.187▲ (45.3%)

REFERENCES

[1] O. Chapelle and Y. Zhang. A dynamic Bayesian network click model for web search ranking. In *WWW*, 2009.

[2] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *TOIS*, 2012.

[3] D. Chen, W. Chen, and H. Wang. Beyond ten blue links: Enabling user click modeling in federated web search. In *WSDM*, 2012.

[4] A. Chuklin, A. Schuth, K. Hofmann, P. Serdyukov, and M. de Rijke. Evaluating aggregated search using interleaving. In *CIKM*, 2013.

[5] A. Chuklin, P. Serdyukov, and M. de Rijke. Using intent information to model user behavior in diversified search. In *ECIR*, 2013.

[6] G. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *SIGIR*, 2008.

[7] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, 2005.

[8] D. Nguyen, T. Demeester, D. Trieschnigg, and D. Hiemstra. Federated search in the wild: the combined power of over a hundred search engines. In *CIKM*, 2012.

[9] T. Sakai. Evaluation with informational and navigational intents. In *WWW*, 2012.

[10] A. Schuth, K. Hofmann, S. Whiteson, and M. de Rijke. Lerot: an online learning to rank framework. In *Living Labs for Information Retrieval Evaluation workshop at CIKM*, 2013.

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

[12] Y. Zhang, L. A. F. Park, and A. Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13(1), 2010.

[13] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating aggregated search pages. In *SIGIR*, 2012.

[14] K. Zhou, M. Lalmas, T. Sakai, R. Cummins, and J. M. Jose. On the reliability and intuitiveness of aggregated search metrics. In *CIKM*, 2013.