Evaluating intuitiveness of vertical-aware click models

Chuklin, A.; Zhou, K.; Schuth, A.; Sietsma, F.; de Rijke, M.

DOI
10.1145/2600428.2609513

Publication date
2014

Document Version
Final published version

Published in
SIGIR '14

Citation for published version (APA):
https://doi.org/10.1145/2600428.2609513
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., [3]). 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

The code we use to simulate clicks is available as a part of Leron [10] at https://bitbucket.org/ilps/lerot
Algorithm 1: Computing the intuitiveness scores of click models $CM_1$ and $CM_2$ based on preference agreement with a gold standard metric $M_{GS}$.

```plaintext
Disagree = 0; Correct1 = 0; Correct2 = 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_2}(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 $(\delta_1 \times \delta_{GS} < 0) \# CM_1$ and $CM_2$ disagree
            Disagree++;
        else
            if $\delta_1 \times \delta_{GS} \geq 0 \# CM_1$ and $M_{GS}$ agree
                Correct1++;
            else
                if $\delta_2 \times \delta_{GS} \geq 0 \# CM_2$ and $M_{GS}$ agree
                    Correct2++;

Intuitiveness($CM_1, CM_2, M_{GS}$) = Correct1/Disagree;
Intuitiveness($CM_2, CM_1, M_{GS}$) = Correct2/Disagree;
```

of clicks they receive according to a click model $CM$, like in A/B-testing experiments. 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 [1, 4]. 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, 4], 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): $P(C_i = 1) = P(C_i = 1 | E_i = 1) \cdot P(E_i = 1)$, where the examination probability $P(E_i = 1) = p^{i-1}$ and the probability of a click given examination is approximated using relevance labels. 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 $\phi_i$ depends only on the rank of the document $i$:

$$P(E_i = 1 | A) = \phi_i = a_{i-1}$$

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 $\gamma_j$ that depends solely on the vertical type $j$:

$$\beta_i(A) = 0$$

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

$$\beta_i(A) = \min\left(1, \frac{1}{\max_{i \cdot A_j = 1} \left| \frac{1}{dist_j(t)} + \gamma_j \right|} \right)$$

(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 [3], 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 [3]. If we do assume that $\gamma_j$ takes different values for different $j$, we get the model that we call here mFCM-NO.

In order to further distinguish different verticals we use the vertical orientation of the user, the probability that users prefer certain verticals 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:

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

(5)

Alternatively, one can perform an interleaving comparison [2], but this is beyond the scope of the paper.
The model defined by equations (1)–(5) is called mFCM. The simpler mFCM-N0 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 $\gamma$, $\phi$ and hpos parameters similar to [4]. We set $\gamma$ 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 [1] for the corresponding vertical types. We also set hpos = [0.95, 0.9, 0.85, 0.8, 0.75, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1] for multimedia verticals as in [4] (assuming the user cannot see documents below rank 6 without scrolling), and for text verticals hpos = [0.95, 0.9, 0.85, 0.8, 0.75, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.0] since Chen et al. [3] suggest that a 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], $\phi$ equals [0.8, 0.6, 0.48, 0.34, 0.28, 0.2, 0.1, 0.08, 0.06] 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 $C_{36}^5 = 630$ run pairs and $50 \times 630 = 31,500$ 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-N0, FCM, PBM, RCM, all of which are described in Section 5. Table I 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 I(a) shows that if we compare mFCM and mFCM-N0 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-N0 is only 0.833, so mFCM is more likely to agree with VS metric than mFCM-N0. Note that the scores of the 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 I(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-N0, FCM;
- VD: $mFCM, mFCM-N0 > PBM > FCM > RCM;
- IS: $mFCM > mFCM-N0 > FCM > PBM > RCM;
- RP: $FCM > PBM > mFCM, mFCM-N0 > RCM;
- VS and IS: $mFCM > mFCM-N0 > FCM > PBM > RCM;
- VS, IS, VD: $mFCM > mFCM-N0 > FCM > PBM > RCM;
- VS, IS, RP, VD (all four metrics): $mFCM > PBM > RCM, FCM > mFCM-N0 > PBM > RCM$.

For single-component evaluation, mFCM outperforms the other models on VS, VD and IS, with mFCM-N0 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-N0 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-N0 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-N0 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, H0R-11-10, 612.066.930, 640.006.013, the Center for Creation, Content and Technology (C3CT), the QuaMities project funded by the CLARIN-nl program, the TBOVE 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 △ (α = 0.05) and ▲ (α = 0.01).

<table>
<thead>
<tr>
<th>Evaluation Criteria</th>
<th>mFCM - NO</th>
<th>FCM</th>
<th>PBM</th>
<th>RCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a. VS) gold standard: vertical selection precision</td>
<td>0.878/0.833△ (14.7%)</td>
<td>0.865/0.842△ (18.2%)</td>
<td>0.873/0.856</td>
<td>0.851/0.865 (44.5%)</td>
</tr>
<tr>
<td>mFCM - NO</td>
<td>0.843/0.850 (17.0%)</td>
<td>0.860/0.877 (21.3%)</td>
<td>0.875/0.885 (22.0%)</td>
<td>0.849/0.871 (44.2%)</td>
</tr>
<tr>
<td>FCM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.846/0.868 (45.3%)</td>
</tr>
<tr>
<td>PBM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(b. VD) gold standard: vertical recall</td>
<td>0.819/0.812 (14.7%)</td>
<td>0.870/0.717 ▲ (18.2%)</td>
<td>0.832/0.748 ▲ (21.5%)</td>
<td>0.860/0.678 ▲ (44.5%)</td>
</tr>
<tr>
<td>mFCM - NO</td>
<td>0.884/0.715 ▲ (17.0%)</td>
<td>0.817/0.743 ▲ (23.1%)</td>
<td>0.763/0.819 ▼ (22.0%)</td>
<td>0.822/0.708 ▲ (44.2%)</td>
</tr>
<tr>
<td>FCM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.828/0.688 ▲ (45.3%)</td>
</tr>
<tr>
<td>PBM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(c. IS) gold standard: mean precision of vertical retrieved items</td>
<td>0.765/0.732△ (14.7%)</td>
<td>0.754/0.691 ▲ (18.2%)</td>
<td>0.822/0.515 ▲ (21.5%)</td>
<td>0.918/0.349 ▲ (44.5%)</td>
</tr>
<tr>
<td>mFCM - NO</td>
<td>0.745/0.706 ▲ (17.0%)</td>
<td>0.815/0.542 ▲ (23.1%)</td>
<td>0.805/0.549 (22.0%)</td>
<td>0.902/0.357 ▲ (44.2%)</td>
</tr>
<tr>
<td>FCM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.828/0.423 ▲ (45.3%)</td>
</tr>
<tr>
<td>PBM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(d. RP) gold standard: Spearman Correlation with &quot;perfect&quot; aggregated search page</td>
<td>0.601/0.592 (14.7%)</td>
<td>0.493/0.691 ▼ (18.2%)</td>
<td>0.575/0.649 ▼ (21.5%)</td>
<td>0.634/0.551 ▲ (44.5%)</td>
</tr>
<tr>
<td>mFCM - NO</td>
<td>0.477/0.702 ▼ (17.0%)</td>
<td>0.569/0.644 ▼ (23.1%)</td>
<td>0.653/0.576 ▲ (22.0%)</td>
<td>0.683/0.513 ▲ (44.2%)</td>
</tr>
<tr>
<td>FCM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.650/0.527 ▲ (45.3%)</td>
</tr>
<tr>
<td>PBM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(e. VS + IS) gold standard: vertical selection precision AND vertical item mean precision</td>
<td>0.666/0.606△ (14.7%)</td>
<td>0.651/0.584△ (18.2%)</td>
<td>0.728/0.432 ▲ (21.5%)</td>
<td>0.782/0.294 ▲ (44.5%)</td>
</tr>
<tr>
<td>mFCM - NO</td>
<td>0.630/0.608△ (17.0%)</td>
<td>0.709/0.474 ▲ (23.1%)</td>
<td>0.714/0.486 ▲ (22.0%)</td>
<td>0.766/0.303 ▲ (44.2%)</td>
</tr>
<tr>
<td>FCM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.700/0.362 ▲ (45.3%)</td>
</tr>
<tr>
<td>PBM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(f. VS + IS + VD) gold standard: vertical selection precision AND vertical item mean precision AND vertical recall</td>
<td>0.567/0.522△ (14.7%)</td>
<td>0.585/0.456△ (18.2%)</td>
<td>0.605/0.347 ▲ (21.5%)</td>
<td>0.680/0.248 ▲ (44.5%)</td>
</tr>
<tr>
<td>mFCM - NO</td>
<td>0.560/0.469 ▲ (17.0%)</td>
<td>0.580/0.372 ▲ (23.1%)</td>
<td>0.568/0.430 ▲ (22.0%)</td>
<td>0.649/0.269 ▲ (44.2%)</td>
</tr>
<tr>
<td>FCM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.601/0.300 ▲ (45.3%)</td>
</tr>
<tr>
<td>PBM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(g. VS + IS + RP + VD) gold standard: ALL single-component metrics</td>
<td>0.372/0.373 (14.7%)</td>
<td>0.346/0.366 (18.2%)</td>
<td>0.394/0.257 ▲ (21.5%)</td>
<td>0.485/0.164 ▲ (44.5%)</td>
</tr>
<tr>
<td>mFCM - NO</td>
<td>0.350/0.370 ▼ (16.0%)</td>
<td>0.390/0.263 ▲ (23.1%)</td>
<td>0.414/0.269 ▲ (22.0%)</td>
<td>0.486/0.155 ▲ (44.2%)</td>
</tr>
<tr>
<td>FCM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.438/0.187 ▲ (45.3%)</td>
</tr>
<tr>
<td>PBM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

REFERENCES