Vertical-Aware Click Model-Based Effectiveness Metrics

Ilya Markov*† i.markov@uva.nl
Eugene Kharitonov‡ kharitonov@yandex-team.ru
Vadim Nikulin‡ vnik@yandex-team.ru
Pavel Serdyukov‡ pavser@yandex-team.ru
Maarten de Rijke† derijke@uva.nl
Fabio Crestani§ fabio.crestani@usi.ch

†University of Amsterdam, Amsterdam, The Netherlands
‡Yandex, Moscow, Russia
§University of Lugano (USI), Lugano, Switzerland

ABSTRACT

Today's web search systems present users with heterogeneous information coming from sources of different types, also known as verticals. Evaluating such systems is an important but complex task, which is still far from being solved. In this paper we examine the hypothesis that the use of models that capture user search behavior on heterogeneous result pages helps to improve the quality of offline metrics. We propose two vertical-aware metrics based on user click models for federated search and evaluate them using query logs of the Yandex search engine. We show that depending on the type of vertical, the proposed metrics have higher correlation with online user behavior than other state-of-the-art techniques.

Categories and Subject Descriptors

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

Keywords

Click models; evaluation; aggregated search

1. INTRODUCTION

When evaluating a web search system, it is commonly assumed that users are presented with ten result snippets, known as "ten blues links," and that these snippets are examined by users from top to bottom. However, current web search systems go beyond the "ten blue links" paradigm and present users with heterogeneous information coming from multiple search engines, also known as *verticals* (e.g., images, news, maps, etc.). In this case user search behavior deviates significantly from that observed in standard web search [3, 10]. Although the changes in user behavior should be taken into account when evaluating heterogeneous search systems, still little is done in this direction [12].

The quality of web search results can be evaluated in two ways: online or offline. Online evaluation, such as A/B-testing or interleaving, gathers feedback directly from users. The feedback usually includes clicks, page dwell times, mouse movements, etc. The quality of a system is then inferred from these signals. Alternatively, web search can be evaluated offline by gathering manual assessments of the quality of a whole search engine result page (SERP) and/or its parts. These assessments may be used directly or within offline effectiveness metrics. Recently, a mixed evaluation approach was proposed, where offline metrics are built based on models of user search behavior and their parameters are learned from query logs [4]. This way evaluation is done offline, producing instant results. Still, it uses direct user feedback (in the form of clicks), thus considering the preferences of actual users.

In this paper we approach the problem of evaluating heterogeneous web search systems by following the above idea. In particular, we develop click model-based effectiveness metrics that account for the presence of a vertical result on a SERP. The research question we address is the following: *is it possible to improve the quality of offline effectiveness metrics for web search by considering user behavior in the presence of a vertical result?*

The contributions of this paper are twofold. First, we develop two vertical-aware effectiveness metrics based on click models for federated search [3, 10]. Second, we evaluate the proposed metrics using a large query log, considering search sessions with different types of vertical results, namely images, video, locations and news.

2. CLICK MODEL-BASED METRICS

Effectiveness metrics in web search are intended to reflect the way users perceive the quality of search results. Increasingly, metrics tend to rely on models of user behavior. Traditional metrics, such as Precision at k or Average Precision, assume that users are interested in relevant documents and, therefore, focus on topical relevance. In addition to relevance, more advanced metrics, e.g., nDCG [7] and RBP [8], assume that users scan results from top to bottom and so discount the relevance of a document by its rank.

Recently, a number of effectiveness metrics were proposed based on user click models. Such models estimate the probability of a click for each document presented to a user. Model-based effectiveness metrics, in turn, use these estimated probabilities to measure the quality of search results. The Expected Reciprocal Rank metric (ERR) [2] uses a simplified version of the DBN click model [1], where a user scans search results from top to bottom until she finds a relevant document or abandons the search. The Expected Browsing Utility (EBU) [11] is also based on the simplified DBN model but, as opposed to ERR, which uses predefined parameter values, EBU estimates parameters directly from click logs.

Chuklin et al. [4] proposed a general way of converting click

*Research mainly carried out while at Yandex, Moscow.

CIKM'14, November 3–7, 2014, Shanghai, China.

ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2661944 .

models into model-based effectiveness metrics and applied this idea to existing models for web search, such as DBN [1], DCM [6] and UBM [5]. As a result, a set of utility- and effort-based metrics were proposed, all of which showed higher correlation with online experiments than the baseline metrics that are not model-based.

3. VERTICAL-AWARE METRICS

The above metrics were shown to be effective in the standard web search scenario. However, existing offline metrics for web search do not consider the presence of vertical results on a SERP. Recent studies showed that user behavior changes considerably in this case [3, 10]. Several click models capturing these changes were proposed—the Federated Click Model (FCM) [3] and the Vertical-aware Click Model (VCM) [10]—showing higher likelihood and lower perplexity compared to click models for standard web search. However, no corresponding effectiveness metrics were developed. We fill this gap by converting the FCM and VCM models into corresponding model-based offline effectiveness metrics. We hypothesize that these metrics correlate better with online experiments than existing offline metrics when a vertical result is present on a SERP.

Both FCM and VCM extend the Utility Browsing Model (UBM) for web search [5] (although DCM- and DBN-based implementations are also possible). Following [4], UBM can be converted into a utility-based metric, so we focus on such metrics in this paper.

A utility-based metric is defined a follows:

$$uMetric = \sum_{k=1}^{N} P(C_k = 1) \cdot r_k \tag{1}$$

where N is the number of documents on a result page, $P(C_k = 1)$ is the probability of the k-th document being clicked and r_k is the relevance of the k-th document.

In Eq. (1) the relevance r_k is derived from relevance judgements created offline, while the click probability $P(C_k = 1)$ is calculated based on a user click model. We follow the work on ERR [2] and define the relevance r based on a relevance grade R as follows: $r = (2^R - 1)/2^{R_{max}}$.

According to the UBM click model, a document is clicked if and only if it is examined and attractive:

$$P(C = 1) = P(E = 1)P(A = 1) = \gamma_{kd}\alpha_{uq},$$

where E and A are random variables denoting examination and attractiveness events. In UBM, attractiveness depends on a document u and query q and the examination probability depends on the position of the document and the distance from the last click.

During offline evaluation of web search, clicks are not available, so the distance d from the last-clicked document is not observed. Therefore, this distance has to be marginalized out in order to calculate the final click probabilities. Following [4], $P_{UBM}(C = 1)$ can be defined recursively as follows:

$$\begin{aligned} P(C_0 = 1) &= 1 \\ P(C_k = 1) &= \sum_{i=0}^{k-1} P(C_i = 1) \left(\prod_{j=i+1}^{k-1} (1 - \gamma_j(i)\alpha_j) \right) \gamma_k(i)\alpha_k \end{aligned} \tag{2}$$

where, for simplicity, $\alpha_k = \alpha_{u_k q}$ and $\gamma_{k(k-i)} = \gamma_k(i)$.

FCM-based Metric.

Studies on user behavior in federated search show that the presence of a vertical result affects examination probabilities of other documents on a SERP [3, 10]. To model this bias, FCM introduces an additional hidden variable F, which indicates whether user behavior changes due to the presence of a vertical result. We will call it *vertical attractiveness* in this paper. The examination probability in FCM is then modeled as follows:

$$\begin{aligned} P(F = 1) &= \phi_{tv} \\ P(E = 1|F = 0) &= \gamma_{kd} \\ P(E = 1|F = 1) &= \gamma_{kd} + (1 - \gamma_{kd})\beta_l, \end{aligned}$$

where t is a type of a vertical result, v is its position and l is the distance between a vertical and other documents on a SERP, which can be both positive and negative. The total examination probability of the FCM model can be calculated as follows:

$$P_{FCM}(E = 1) = \gamma_{kd} + (1 - \gamma_{kd})\phi_{tv}\beta_l.$$

In order to obtain the click probability $P_{FCM}(C = 1)$, the examination probability $P_{FCM}(E = 1)$ should be plugged into Eq. (2) instead of $P_{UBM}(E = 1) = \gamma_{kd}$. Then, the ***uFCM metric*** can be constructed by plugging $P_{FCM}(C = 1)$ into Eq. (1).

VCM-based Metric.

Similarly to FCM, VCM assumes that examination probabilities change when an attractive vertical result is present on a SERP ($F = 1$). Additionally, VCM assumes that in this case a user examines the vertical result first and then continues examining other results in either a bottom-up or top-down direction. This is controlled by a hidden variable B. Overall, VCM models the examination probability as follows:

$$\begin{aligned} P(F = 1) &= \phi_{tv} \\ P(B = 1|F = 0) &= 0 \\ P(B = 1|F = 1) &= \sigma_{tv} \\ P(E = 1|F = 0) &= \gamma_{kd} \\ P(E = 1|F = 1) &= \gamma'_{kd}. \end{aligned}$$

These equations define three possible examination trails for a SERP: (i) starting from the top document down to the bottom ($F = 0$), (ii) starting from a vertical and then back to the top ($F = 1, B = 1$), and (iii) starting from a vertical down to the bottom ($F = 1, B = 0$). The total examination probability in VCM is the weighted average of examination probabilities over these trails:

$$P_{VCM}(E = 1) = (1 - \phi_{tv})\gamma_{kd} + \phi_{tv}\sigma_{tv}\gamma'_{kd'} + \phi_{tv}(1 - \sigma_{tv})\gamma'_{kd''}$$

where d, d' and d'' are the distances from the last clicked document according to each trail.

The overall examination probability of the VCM model cannot be directly plugged into Eq. (2), because it uses different distances for different trails. Each distance should be marginalized separately by deriving a click probability for each trail. Then the overall click probability of the VCM model can be calculated as follows:

$$\begin{aligned} P_{VCM}(C = 1) = (1 - \phi_{tv})P_1(C = 1) + \phi_{tv}\sigma_{tv}P_2(C = 1) \\ + \phi_{tv}(1 - \sigma_{tv})P_3(C = 1), \end{aligned}$$

where P_i is the click probability for i-th trail. The ***uVCM metric*** is obtained by plugging $P_{VCM}(C = 1)$ into Equation (1).

4. EVALUATION

4.1 Experimental Setup

In order to evaluate the proposed vertical-aware effectiveness metrics, we collected user search sessions from click logs of a large commercial search engine, namely Yandex. Similarly to [3, 10], we use vertical results of three different types: images and video represent multimedia verticals, news comprises a text-based vertical

Table 1: Summary of user search sessions in training and test sets for different verticals.

Training sets			
vertical	# queries	# configurations	# sessions
images	21,432	211,717	313,138
video	6,989	70,697	111,417
locations	745	20,532	71,753
news	352	1,407	1,994
Test sets			
vertical	# queries	# configurations	# sessions
images	21,316	210,622	311,384
video	7,062	60,352	95,489
locations	733	20,378	71,456
news	424	1,516	2,069

and locations represent a vertical with composite results, containing both textual and visual information. We sampled sessions, containing one of these vertical results, during November 2013. The top-10 documents in each session were judged by human assessors using the standard five-grade scale (perfect, excellent, good, fair, and poor). The collected sessions are split based on user ids into training and test sets of roughly the same size (see Table 1). The uneven distribution of the number of sessions between verticals is due to the difference in frequency with which vertical results were triggered in the sampled part of our click logs.

Following [2, 4], we evaluate the quality of the proposed metrics based on their correlation with online metrics, such as UCTR and Max/Mean/MinRR. UCTR is a binary variable showing if there was a click in a session or not (opposite to abandonment). MeanRR is the mean reciprocal rank of clicks in a session, MaxRR is the reciprocal rank of the first click and MinRR is the reciprocal rank of the last click. For the above online metrics only clicks on web results are considered.

Since for the same query a SERP may vary depending on a user, her location, etc., we focus on *configurations* [2], which is a query with a fixed SERP (see Table 1 for statistics). Offline metrics produce the same value for the same configurations, while the values of online metrics are averaged over all sessions with the same configuration. The weighted correlation between offline and online metrics is calculated over all configurations as in [2]:

$$Corr = \frac{\sum_{c=1}^{N} n_c(m_1(c) - \bar{m}_1)(m_2(c) - \bar{m}_2)}{\sqrt{\sum_{c=1}^{N} n_c(m_1(c) - \bar{m}_1)^2}\sqrt{\sum_{c=1}^{N} n_c(m_2(c) - \bar{m}_2)^2}},$$

where N is the total number of configurations, n_c is the number of occurrences of the configuration c, $m_i(c)$ is the value of the metric m_i for the configuration c and $\bar{m}_i$ is the mean value of m_i.

We compare our vertical-aware metrics against two types of baseline: (i) static offline metrics where parameters are fixed (DCG and ERR), and (ii) click model-based metrics for web search, where parameters are learned from click logs (EBU, uDCM, uDBN and uUBM). When learning model parameters, the attractiveness probability $P(A = 1)$ (and the satisfaction probability $P(S = 1)$ for DBN) is assumed to be dependent only on the relevance grade of a document given a query as in [4].

4.2 Results and Discussion

The weighted correlation between offline and online metrics for different types of vertical results is shown in Tables 2–5, the best values are given in bold. Table 2 presents results for the news vertical. News snippets contain mainly text and are, therefore, similar to standard web snippets. Due to this, most offline metrics (apart from DCG) have similar correlation with online metrics. Still, the proposed vertical-aware metrics, uFCM and uVCM, are slightly superior to others.

Table 2: Weighted correlation between offline and online metrics when a news vertical result is present on a SERP.

	MaxRR	MeanRR	MinRR	UCTR
DCG	0.2390	0.3082	0.3481	0.1794
ERR	0.2562	0.3306	0.3732	0.1864
EBU	0.2588	0.3324	0.3748	0.1967
uDCM	0.2569	0.3304	0.3728	0.1952
uDBN	0.2682	0.3429	0.3856	0.2034
uUBM	0.2669	0.3421	0.3851	0.2012
uFCM	**0.2703**	0.3456	0.3886	**0.2044**
uVCM	0.2702	**0.3459**	**0.3892**	0.2033

Table 3: Weighted correlation between offline and online metrics when an image vertical result is present on a SERP.

	MaxRR	MeanRR	MinRR	UCTR
DCG	0.1979	0.2394	0.2559	0.1526
ERR	0.2170	0.2634	0.2823	0.1554
EBU	0.2110	0.2581	0.2774	0.1551
uDCM	0.2090	0.2563	0.2759	0.1529
uDBN	0.2216	0.2704	0.2905	0.1599
uUBM	0.2184	0.2672	0.2875	0.1566
uFCM	**0.2495**	**0.2973**	**0.3144**	**0.1917**
uVCM	0.2222	0.2713	0.2914	0.1615

Tables 3 and 4 present results for multimedia verticals, namely images and video. In both cases, uFCM achieves much higher correlation values with all online metrics than the baselines. This result is intuitive considering that user behavior was reported to change considerably when a visually attractive vertical result (e.g., an image) is present on a SERP [3, 10]. The FCM model captures such changes, which, in turn, results in higher correlation values between uFCM and online metrics.

The uVCM metric is the second best among model-based metrics in terms of correlation values with online experiments, but it does not correlate as well as uFCM. This can be explained as follows. The FCM and VCM click models both use the vertical attractiveness parameter ϕ_{tv}, which shows how much user behavior deviates from the standard web search scenario when a vertical result of type t is shown at rank v. The lower the value of ϕ_{tv}, the closer a vertical-aware model is to the underlying UBM model. After training FCM and VCM for the image and video verticals, we observed that FCM estimated ϕ_{tv} to be relatively high, which means that FCM deviates considerably from UBM; in contrast, VCM estimated ϕ_{tv} to be quite low, thus being close to UBM. Indeed, Tables 3 and 4 show that the correlation of uVCM with online metrics is somewhat close to that of uUBM.

Table 5 presents results for the location vertical, which consists of both textual and visual information. Interestingly, DCG has the highest correlation with RR-based online metrics, followed by uDCM (which has the highest correlation with UCTR) and EBU. In order to get insights into these results, we conducted an A/B-testing

Table 4: Weighted correlation between offline and online metrics when a video vertical result is present on a SERP.

	MaxRR	MeanRR	MinRR	UCTR
DCG	0.1850	0.1982	0.1960	0.1538
ERR	0.2709	0.2876	0.2849	0.2424
EBU	0.2050	0.2185	0.2157	0.1681
uDCM	0.2037	0.2168	0.2137	0.1668
uDBN	0.2518	0.2663	0.2624	0.2171
uUBM	0.2465	0.2608	0.2568	0.2111
uFCM	**0.3034**	**0.3155**	**0.3074**	**0.2704**
uVCM	0.2611	0.2753	0.2708	0.2259

Table 5: Weighted correlation between offline and online metrics when a location vertical result is present on a SERP.

	MaxRR	MeanRR	MinRR	UCTR
DCG	**0.2107**	**0.2288**	**0.2380**	0.1321
ERR	0.1606	0.1772	0.1858	0.0872
EBU	0.1957	0.2093	0.2147	0.1372
uDCM	0.1966	0.2103	0.2158	**0.1379**
uDBN	0.1887	0.2046	0.2122	0.1175
uUBM	0.1887	0.2052	0.2134	0.1163
uFCM	0.1804	0.1949	0.2014	0.1156
uVCM	0.1816	0.1983	0.2067	0.1105

experiment on real users of the considered search engine, where the location vertical was suppressed for a period of one week. The experiment showed that the abandonment rate of the control (the vertical result is displayed) was significantly higher than for the treatment (the vertical result is suppressed). There might be two reasons for this: (i) users are satisfied with the information presented on a SERP (address, phone number, working hours, etc.) and leave the search without any click, which is known as good abandonment, or (ii) some users consider web results above the location vertical as a banner (especially for high positions of the vertical) and skip them, which is known as banner blindness. For navigational queries this results in no clicks on a SERP. In both cases, online metrics, like MeanRR and UCTR, do not fully capture the underlying user behavior. Thus, the low correlation of offline metrics in Table 5 cannot be interpreted as a failure. Instead, other means of evaluating the quality of offline metrics must be used (e.g., classifying abandonments into "good" and "bad" as in [9] and calculating correlation only for the latter), which we plan to do as future work.

Overall, our results show several important trends. First, they confirm the findings of previous studies on user behavior in federated search, that is, user behavior depends on the type of a vertical result present on a SERP, where visually attractive verticals, e.g., video, affect this behavior more than text-based ones such as news. Mixed-content verticals, such as the location vertical, trigger more complex user behavior, which requires further investigation.

Second, in answer to the research question posed in Section 1, we showed that, depending on the type of vertical, the proposed vertical-aware click model-based metrics have higher correlation values with online user behavior than other offline metrics for web search. In particular, uFCM has the highest correlation when a visualy attractive vertical, i.e. image or video, is present on a SERP. The uVCM metric, on the other hand, is more conservative, being closer to the underlying UBM model.

5. CONCLUSIONS AND FUTURE WORK

In this paper we approached the problem of offline evaluation for heterogeneous web search environments, where standard web results are augmented with results from vertical search engines. We investigated whether considering user behavior on such federated SERPs helps to improve the quality of offline metrics. To this end, we considered existing click models for federated search, namely FCM and VCM, and converted them into click model-based effectiveness metrics. Experimental results showed that, depending on the type of vertical, the proposed metrics have higher correlation values with online metrics, and especially so when visually attractive vertical results, such as images or video, are present on a SERP.

As future work, we plan to extend the proposed metrics to evaluate not only web results, but a SERP as a whole, including vertical results, sponsored search and other components. We also plan to investigate user behavior when a location vertical result is present on a SERP in more detail. We first need to understand the cause of the high abandonment rate observed in this case and then we plan to learn to distinguish between good and bad abandonments for a more precise evaluation of the quality of offline metrics.

Acknowledgments. The authors would like to thank Eugene Krokhalev and Sergey Protasov for inspiring discussions and technical support. This research was partially funded by grant P2T1P2_152269 of the Swiss National Science Foundation, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 288024 (LiMoSINe) and nr 312827 (VOX-Pol), the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, the Center for Creation, Content and Technology (CCCT), 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.

6. REFERENCES

[1] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *WWW '09*, pages 1–10, 2009.
[2] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM '09*, pages 621–630, 2009.
[3] D. Chen, W. Chen, H. Wang, Z. Chen, and Q. Yang. Beyond ten blue links: enabling user click modeling in federated web search. In *WSDM '12*, pages 463–472, 2012.
[4] A. Chuklin, P. Serdyukov, and M. de Rijke. Click model-based information retrieval metrics. In *SIGIR '13*, pages 493–502, 2013.
[5] G. E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *SIGIR '08*, pages 331–338, 2008.
[6] F. Guo, C. Liu, and Y. M. Wang. Efficient multiple-click models in web search. In *WSDM '09*, pages 124–131, 2009.
[7] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Information Systems*, 20(4):422–446, 2002.
[8] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Information Systems*, 27(1):2:1–2:27, 2008.
[9] Y. Song, X. Shi, R. W. White, and A. Hassan. Context-aware web search abandonment prediction. In *SIGIR '14*, 2014.
[10] C. Wang, Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, and K. Zhang. Incorporating vertical results into search click models. In *SIGIR '13*, pages 503–512, 2013.
[11] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Expected browsing utility for web search evaluation. In *CIKM '10*, pages 1561–1564, 2010.
[12] K. Zhou, T. Sakai, M. Lalmas, Z. Dou, and J. M. Jose. Evaluating heterogeneous information access. In *Proc. MUBE workshop*, 2013.