Web Search Engine Evaluation Using Clickthrough Data and a User Model

Georges Dupret
Yahoo! Research
Latin America
gdupret@yahoo-inc.com

Vanessa Murdock
Yahoo! Research Barcelona
vmurdock@yahoo-inc.com

Benjamin Piwowarski
Yahoo! Research
Latin America
bpiwowar@yahoo-inc.com

ABSTRACT

Traditional search engine evaluation relies on a list of query document pairs along with a score reflecting the document relevance to the query. The score is generally a human assessment, but nothing is said explicitly about the actual user behavior. In this paper we illustrate with a toy model that once the user behavior is agreed upon, the human assessment can be eliminated and the engine performance can be evaluated based on the clickthrough data of past users.

1. INTRODUCTION

The majority of research in evaluating information retrieval systems has focused on static collections consisting of relatively homogeneous documents, and queries composed or selected by professional assessors, such as the collections assembled by NIST[1]. Even in these controlled conditions, constructing assessments is an expensive and time consuming activity. Evaluating a Web search engine is considerably more resource-consuming as the Web is dynamic rather than static, the documents are heterogeneous in terms of quality, content, format and language, and the queries are posed by the general public, rather than information finding experts. That said, Web search engines have a potentially unlimited source of information about documents and queries that could not possibly be constructed in a laboratory setting: the complete user session (query history, viewed documents, clicked documents, etc.) for millions of users, in millions of search contexts. If we could create a model of the user interaction with the system in such a way that data about clicked documents could be used as a surrogate for relevance judgments, search engines could be evaluated more accurately, improving the quality of the results with very little cost or overhead. Furthermore, an automated Web search engine evaluation is able to keep pace with document collection changes, or changes in the focus of the popular searches.

A difficulty arises in the use of user clicks as relevance surrogates. Although it is possible to analyze millions of sessions, a user click has been found to be a weak indicator of user interest [11]. One reason for this is that although most users scan the results list from the top, each user halts their search at a different position in the ranked list. Thus, we know how many times a document was selected, but we don't know how many times it was viewed by the user who decided not to select it. Furthermore, documents presented in the first page of results are more likely to be clicked than documents presented in later pages, and documents presented at the top of the ranked list are more likely to be clicked than documents presented further down, irrespective of their relevance [8]. Several methods have been proposed to compensate for this bias [3, 13].

In this paper we propose to evaluate a Web search engine, we must estimate the probability that a document is relevant to a set of queries, unbiased with respect to the document's position in the ranked list. Further, we must estimate the probability that a user will view the document snippet in the ranked list. Intuitively, the user frequently views only a portion of the ranked list, and documents for which the URL, title and text snippet are never seen will not be clicked.

We propose a generative model to predict user clicks on document snippets based on user sessions recorded in the query logs. Our model unbiases the clicks with respect to position in the ranked list, allowing the data to be used as relevance surrogates to evaluate search engine performance.

The remainder of the paper is organized as follows. In Section 2, we formalize our hypotheses and present a summary of the model. We present experiments in Section 3 to investigate the properties of the model itself, compared to a simple model where documents are selected irrespective of their positions. In this section we propose a metric for search engine evaluation based on our model, and examine the stability of the measure. We discuss related work and present our conclusions and future work in Section 4.

2. A GENERATIVE MODEL

Consider the following scenario: A user issues a query, and is presented with a list of document snippets by the search engine. The user scans the list starting with the first result and goes down the list of document snippets sequentially, according to their rank. For each snippet in the list, the user either selects the document because the snippet was attractive, or the snippet was not selected. (We do not know whether the snippet was not selected because it was unattractive, or because the user did not look at it.) The user then returns to the list and either continues the search or abandons it. Although real user behavior may be considerably more complex, for example, a user may go back to previous results in the list, or jump ahead in the list without looking at the intervening documents, this scenario is consistent with the eye-tracking studies of search behavior described in Joachims et al. [5, 8]. In a sample of more than ten million sessions, we observed that more than 91% con-

[1]http://www.trec.nist.gov (April 2007).

WWW2007, May 8–12, 2007, Banff, Canada.
.

tain only sequential selections. In the following sections we present a generative model that represents this scenario.

2.1 Variables and assumptions

The selection process can be represented as the joint probability $P(s, d, u, q)$ where q is the query (a string), u the document (an URL), d the distance to the previous selection in the same session[2] and s reflects whether the document was selected or not.

Whether the document snippet is selected depends in part on its *attractivity*. This is a binary decision: either the user finds the snippet attractive enough to click on, or not. Thus the probability of the attractivity of a document can be seen as the result of a voting process among all users issuing a given query and who saw the document.

The decision to continue searching up to the position of a given document snippet is the *perseverence*. We propose that perseverence is dependent on the distance from the last selection. The intuition is that a user tends to abandon the search after seeing a long sequence of unattractive snippets.

If we make the assumption that the snippet represents the document fairly, the probability of attractivity can be interpreted as a measure of relevance. This is the approach taken by Radlinski and Joachims [12] among others.

While click-through data is noisy, the user selections do convey information and the effects of occasional user selection mistakes will be mitigated by considering a large number of selections. Presumably, the noisier the data, the larger the number of query sessions needed to infer a relationship between clicks and relevance. The question of how many sessions of a query are needed is an open question, but we assume our Bayesian model copes gracefully with this variability. Moreover, commercial search engines generate enormous amounts of data, and even the comparatively small amount used in experiments reported in the literature [12, 9] produced significant results.

2.2 The Model

If we include the latent variables a (*attractivity*) and c (*consideration*), the full model is identified with the joint distribution $P(s, a, c, u, q, d)$. We view *attractivity* as an intrinsic property of the relation between the document and the query. This is precisely the property the search engine attempts to evaluate in order to rank the documents. We also suppose that the user decision to continue considering snippets after scanning without success through $d-1$ positions happens before he examines the snippet at distance d. Consequently, his decision is independent of the actual snippet content at distance d. Conditioning on u, q and d, the full joint distribution can be rewritten as:

$$P(\mathbf{s}, \mathbf{a}, \mathbf{c}|u, q, d) \quad = \quad P(\mathbf{c}|d)P(\mathbf{a}|u, q) \qquad (1)$$

where we use boldface to denote that a random variable is true (e.g. $\mathbf{s}$). The probability $P(\mathbf{s}|\mathbf{a}, \mathbf{c})$ is deterministic (and equals 1) because a user selects a document only if its snippet is attractive and considered.

Attractivity can be modeled by a simple Bernoulli trial with success probability $\alpha_{u,q}$ because it is a binary value depending exclusively on the document and the query. Consideration is also modeled by a Bernoulli trial, with a success probability γ_d, which we refer to as the *perseverence*, representing the probability of considering a snippet at a distance d knowing that the user did not select any document in the previous $d-1$ positions.

Beta distributions $\text{Be}(\alpha_{uq}|a_{uq}, b_{uq})$ and $\text{Be}(\gamma_d|m, n)$ are adopted as priors for each α_{uq} and γ_d, where a_{uq}, b_{uq}, m_d and n_d are the prior parameters associated with the (u, q) pairs and the perseverences at different distances.

A special case of our model is the "popularity" or "naïve" model where perseverence is constant, i.e. when $P(\mathbf{c}|d) = 1$. In this special case the attractivity of a document is the number of sessions for the query, divided by the number of selections of the document.

We make the common assumption that observations are independent: the data likelihood is the product of the individual observation likelihoods. This simple model can be extended in various ways, for example to take into account the positions of the documents or the page of results. The model and an iterative algorithm for parameters estimation are described in more detail in Dupret et al. [4].

Table 1: Mean log-likelihood per session for the "popularity" and "distance" models for different priors $\text{Be}(\alpha_{uq}|a_{uq}, b_{uq})$ (all documents have the same prior). Reasonable $\text{Be}(\gamma_d|m_d, n_d)$ priors have negligeable influence and we chose $m_d = n_d = 1$ for all d.

a	b	popularity	distance
1	1	-2.85	-2.26
1	10	-2.86	-1.93
1	100	-2.93	-1.90
0.1	10	-2.91	-1.76
1	1000	-3.32	-2.65

3. NUMERICAL EXPERIMENTS

To evaluate the model, we compare it to a popularity model. Following, we develop a metric to determine whether the proposed model is useful for evaluating search engines.

3.1 Model Evaluation

We evaluate the model on a sample of queries from a commercial search engine. There are approximately 3 million (u, q) pairs. The average log-likelihood of the data per session given the naïve model compared to the distance-centric model for different priors $\text{Be}(\alpha_{uq}|a_{uq}, b_{uq})$ is shown in Table 1. The parameters a_{uq} and b_{uq} correspond to a prior of a_{uq} selections in a total of $a_{uq} + b_{uq}$ sessions. For example, $a_{uq} = 1$ and $b_{uq} = 10$ corresponds to 1 selection in 11 sessions. Using ten random test and training splits, we find that the distance model, whatever the value of a_{uq} and b_{uq}, has a higher log-likelihood than the naïve model, with a t-test p-value of $\approx 6 \times 10^{-13}$. We conclude that the distance model better represents the data than the popularity model.

The priors associated with the perseverence are dominated by the data and have only a marginal influence on the results. We set the prior to be $\text{Be}(\gamma_d|1, 1)$ for all d, corresponding to a uniform prior.

The priors associated with attractivity have a strong influence on the final attractivity estimates. The reason is that for most document-query pairs, we have relatively few observations. The larger the value $a_{uq} + b_{uq}$, the more confidence we place in our prior and more observations from the

[2]The meaning of "session" in this work is unsual and represents the set of repetitions of the same query.

logs are necessary to alter it. The choice of prior therefore allows us to tune the system to give more or less confidence in user clicks compared with other sources of information, such as the engine score or the Pagerank of the document. In presence of sparse and noisy data like clickthrough data, this is clearly an advantage.

3.2 Search Engine Evaluation

The main proposition of this paper is that given a user model, click-through data can be used to evaluate a search engine, under the assumption that the attractivity of a document snippet can be used to estimate document relevance. Because the model removes the positional bias in the click-through data, the attractivities do not depend on how the user reached the document, but rather on whether a user who saw the document selected it or not. The engine evaluation measure we propose is the expected number of attractive documents a user would see.

Because search engines regularly update their document collection and their ranking algorithms, the position of a document in the result list may change. We denote a specific ordering by o. Such ordering is typically associated with several sessions and we denote $P(o|q)$ its probability of occurence. We call σ a sequence of selections in a result list, that is the set of ranks the user clicked on. The probability $\mathrm{P}(\sigma|q,o)$ is computed with our model, considering the attractivity of the presented documents and the persistence of the users. For a given set of selections, we can compute the expected number of attractive documents the user sees as $a(\sigma,o,q) = \sum_{r\in\sigma} P(a_r|q,o)$ where a_r is the attractivity of the document at rank r for the ordering o of query q. In these settings, the expected number of attractive documents is

$$\mathcal{R} \quad = \quad \sum_q \mathrm{P}(q) \sum_o \mathrm{P}(o|q) \sum_\sigma \mathrm{P}(\sigma|o,q) a(\sigma,o,q)$$

The $\mathcal{R}$ score is conditioned on the users effectively following the selection process we described. The measure is thus dependent on both the logs (for the queries and the orderings) and the user model. The measure is higher when most probable user behaviors, queries and orderings, as estimated by the $\mathrm{P}(\sigma,o,q)$ term, are related to a high number of attractive documents $a(\sigma,o,q)$.

It may seem counter-intuitive to use the same model to measure performance and to set the model parameters. However, we can state that if the model is accurate enough and the parameters are correctly learned, then the measure is a good approximation of the real average number of attractive documents an average user would see.

The $\mathcal{R}$ score can be understood as generalization of Discounted Cumulative Gain (DCG [10]). According to this last measure, the gain of a document depends on its relevance and on its rank. Intuitively, a relevant document at rank 10 is less useful than a relevant document at rank one, so the relevance score of a document is discounted by the log of its rank. However, whereas cumulative gain was constructed in a standard IR evaluation setting and its paramters are set heuristically, the measure $\mathcal{R}$ directly estimates its parameters according to the query logs. The equivalent of the discount parameter in DCG is the persistence of the user and leads in practice to a decreasing "gain" for higher ranks. In Figure 1 we plotted the mean probability of a document consideration as a function of its position on a log scale.

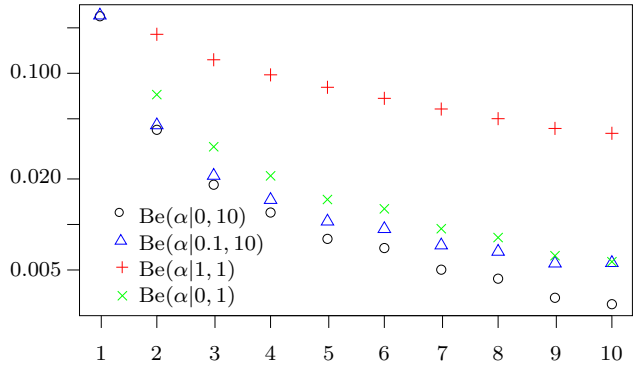

Figure 1: Mean probability of considering the position on the x-axis for various attractivity priors (log scale).

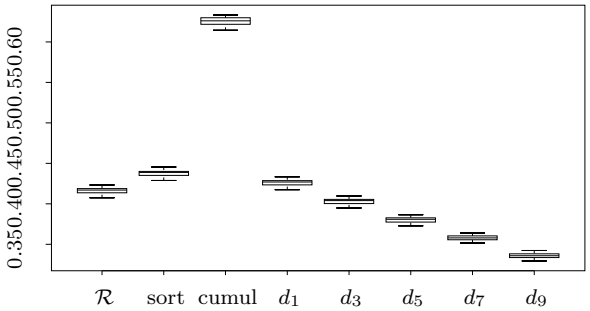

Figure 2: Boxplots of $\mathcal{R}$ scores and various statistics for the prior $\mathrm{Be}(\alpha|0,10)$. The column labeled "sort" shows the score achieved if documents are sorted by decreasing attractivities. "cumul" is the sum of mean attractivities over the first ten ranks, and d_1 to d_9 are obtained by replacing 1 - 9 documents at a random position by a document with attractivity 0.

We observe an exponential decrease similar to the discounting factor in the DCG measure. The $\mathcal{R}$ score shares some of the characteristics of DCG, but it has been adapted to the user behavior of the engine being studied. Similarly, the attractivity is estimated by unbiasing the click-through data. Another difference is that the averaging over queries is done with respect to the query distribution in the logs.

We estimated the $\mathcal{R}$ scores for different priors on attractivities (perseverance priors are always set to $\mathrm{Be}(\gamma_d|1,1)$) and 30 different random splits of the query logs. We split the data by identifying the orderings in the click-through data and we assign them randomly to the training or test set with a probability one half. Consequently, an ordering appearing in the training set never appears in the test set, even if there is more than one query session for a given ordering. We learn the attractivities and perseverences on the training set and use this data to estimate the $\mathcal{R}$ score on the test set. In Table 2, we report the mean $\mathcal{R}$ scores and the associated variance for different priors computed over the first ten positions in the ranking. We see that the variance is remarkably small for all realistic priors, ensuring that the

Table 2: $\mathcal{R}$ (scaled by a constant in order to mask proprietary information) scores and various statistics for different priors Be$(\alpha|a_{uq}, b_{uq})$. We consider only the first 10 positions.

a	b	$\mathcal{R}$	var $\mathcal{R}$	sorted	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	cumul
0	1	0.59	0.10	0.65	0.63	0.61	0.58	0.56	0.54	0.52	0.50	0.48	0.46	1.00
0	10	0.41	0.07	0.43	0.42	0.41	0.40	0.39	0.38	0.36	0.35	0.34	0.33	0.62
0.1	1	0.67	0.08	0.76	0.75	0.74	0.73	0.73	0.72	0.71	0.70	0.70	0.69	1.37
0.1	10	0.43	0.06	0.46	0.45	0.44	0.42	0.41	0.40	0.39	0.38	0.37	0.36	0.69
10	1	1.99	0.30	2.49	2.51	2.53	2.54	2.56	2.58	2.60	2.62	2.64	2.67	8.19
1	1	1.09	0.11	1.41	1.43	1.44	1.46	1.48	1.50	1.52	1.55	1.57	1.60	3.92
1	10	0.55	0.04	0.60	0.59	0.59	0.58	0.58	0.57	0.57	0.56	0.56	0.55	1.28

measure is resistant to changes in the training set.[3] The "cumul" column in Table 2 reports the sum of the document attractivities over the ten positions, averaged over the 30 splits, corresponding to a user selecting all documents.

We also report the performance of an ideal search engine that orders the documents according to decreasing attractivities under the label "sorted". As expected, the score is systematically and significantly higher than the actual engine score. This is naturally the best achievable score for natural settings of the perseverences (it is possible to choose perseverences such that other rankings lead to a better score, but these are unrealistic).

The $d_i, i = 1 \dots 9$ columns report the results of progressively degrading the optimal ranking by randomly choosing i documents among the first ten positions and replacing them with a document having an attractivity distribution equal to the prior. We observe a stable decrease along with the level i of degradation for all but the unreasonable priors, where in most cases the prior attractivity is higher than the one of the replaced documents. This shows that the measure degrades gracefully with the quality of the ranking.

4. RELATED WORK AND DISCUSSION

The majority of work on click-through behavior aims at infering relevance judgments from user clicks. User clicks are exploited to re-rank search results by Joachims [6] and Radlinski and Joachims [12], as well as in Agichtein et al. [1] or [3]. Carterette and Jones use click-through data to predict search engine performance using a probabilistic approach [2]. Joachims et al. [7] recommend pages to users based on their previous selections, and the selections of users who browsed the same information.

The toy model we present is clearly unable to represent accurately user bahavior, but it opens the door to more formal analysis of click-through data. In future work, perseverence could be modeled to depend on the page of results and on the query to reflect that a user may have different behavior for different queries. Other interest indicators, such as time or query history, might be included in this model. Because it is a bayesian model, it can easily be extended to include other types of evidence in the priors. In spite of its deficencies, the current model helped us to argue that given a reasonable user model, a search engine can be evaluated based on the click-through data.

[3] Be$(\alpha_{uq}|10, 1)$ implies that we believe that all documents are 90% attractive beforehand.

5. REFERENCES

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of SIGIR*, 2006.

[2] B. Carterette and R. Jones. Evaluating web search engines using clickthrough data. Submitted to SIGIR 2007.

[3] G. Dupret, B. Piwowarski, C. Hurtado, and M. Mendoza. A statistical model of query log generation. In *Proceedings of the 13th International Symposium on String Processing and Information Retrieval (SPIRE 2006)*, LNCS 4209, pages 217–228. Springer, 2006.

[4] G. Dupret, B. Piwowarski, and V. Murdock. A bayesian model of query log generation. Submitted to SIGIR 2007.

[5] L. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *Proceedings SIGIR*, 2004.

[6] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of KDD*, 2002.

[7] T. Joachims, D. Freitag, and T. Mitchell. WebWatcher: A tour guide for the World Wide Web. In *Proceedings of IJCAI*, pages 770–777, 1997.

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

[9] R. Jones and D. Fain. Query word deletion prediction. In *Proceedings of SIGIR*, 2003.

[10] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[11] D. Kelly. Implicit feedback: Using behavior to infer relevance. In A. Spink and C. Cole, editors, *New Directions in Cognitive Information Retrieval*, pages 169 – 186. Springer Publishing, Netherlands, 2005.

[12] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *Proceedings of KDD*, 2005.

[13] F. Radlinski and T. Joachims. Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 2006.