Sampling High-Quality Clicks from Noisy Click Data

Adish Singla
Bing Search, Microsoft
Richmond, BC, Canada V6V 2J8
adishs@microsoft.com

Ryen W. White
Microsoft Research
Redmond, WA 98052
ryenw@microsoft.com

ABSTRACT
Click data captures many users' document preferences for a query and has been shown to help significantly improve search engine ranking. However, most click data is noisy and of low frequency, with queries associated to documents via only one or a few clicks. This severely limits the usefulness of click data as a ranking signal. Given potentially noisy clicks comprising results with at most one click for a query, how do we extract high-quality clicks that may be useful for ranking? In this poster, we introduce a technique based on query entropy for noise reduction in click data. We study the effect of query entropy and as well as features such as user engagement and the match between the query and the document. Based on query entropy plus other features, we can sample noisy data to 15% of its overall size with 43% query recall and an average increase of 20% in precision for recalled queries.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]**:** Information Search and Retrieval – *search process, retrieval models*

General Terms
Measurement, Experimentation

Keywords
Click data, query entropy, noise elimination, Web search ranking

1. INTRODUCTION
Click data are generated when users query search engines and click on returned search results. These data can be useful for Web search ranking [1,3,6,10], as judgment labels for ranker training [8], query suggestion [4], etc. Click data can be represented as the click counts associated with query-document pairs. However, click data utility is limited by sparseness, low volume, and noise.

Sparseness refers to missing click data for query-document pairs which makes it hard for a ranker to effectively learn features from clicks and challenging to rank all documents based on clicks. Smoothing techniques such as random walk and others handle such sparseness through query and document similarity [6,10]. Documents associated with low volume queries have only a few clicks and hence lack a clear signal about which documents users prefer. Noisy data may contain inconsistencies or randomness in normal user behavior, sometimes associated with low volume.

We address the challenge of sampling high-quality clicks from noisy click data using features of query entropy, user engagement, and relevance. Deng et al. [4] and Dou et al. [5] used entropy for tasks such as query suggestion and measuring click variability for a query. Agichtein et al. [1] used engagement and relevance in the form of query-text overlap to train a ranking algorithm from click data. We also use entropy, but for a different purpose, and study features that may help sample high-quality clicks rather than train ranking algorithms. That said, effectively sampling high-quality clicks may ultimately offer potentially useful ranker training data.

WWW 2010, April 26–30 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

2. CLICK DATA PREPARATION

2.1 Click Data
Click data were generated from anonymized logs of URLs visited by users who opted in to provide data through a widely-distributed browser plugin during a four-month period from February 2009 through April 2009 inclusive. This represented billions of URL visits from millions of users in the English-speaking United States ISO locale. Log entries included a user identifier, a timestamp for each page view, the URL of the Web page visited, and the page title. From these logs, click data of the form: $< q, d, c, fc >$ were extracted, where c is the total number of times a document d is clicked for query q and fc counts only clicks when d was the first result clicked for the query in cases where a user clicked multiple search results. In addition, we extracted a set of other features as discussed in Section 3. This gave us 206 million query-document pairs with 75 million unique queries and 119 million unique documents. 90% of these query-document pairs had only one click, 5% had only two clicks and only 1% had over five clicks.

2.2 Human-Judged Relevance Data
We also obtained human relevance judgments for twelve thousand queries randomly sampled by frequency from the query logs of a large search engine. Trained judges assigned relevance labels from a six-point scale – *Bad*, *Poor*, *Fair*, *Good*, *Excellent*, and *Perfect* – to top-ranked pooled search results for each query from Google, Yahoo! and Bing. We use these judgments to estimate the relevance of documents associated to queries in the click data.

2.3 Scoring and Ranking Approach
Each query-document pair in our data is assigned a score from raw click count using: $Score(q,d) = (1 + fc)/(1 + \sum_d fc)$ .This is a slight variant of last click count**-**based scoring used in [6]. The additional factor of one in the numerator ensures that a document in the click data for a query receives a positive score even if fc is zero (i.e., document never the first result clicked by users for q).

To compute the ranking accuracy of click data, we used a simple approach where we rank documents clicked for a query in descending order of $Score(q,d)$. The ranked lists produced are compared to ideal rankings from relevance judgments. We used Normalized Discounted Cumulative Gain (NDCG) [7] at position one ($NDCG@1$) as a measure of rank precision. NDCG is defined as: $NDCG(i) = N_i \sum_i (2^{r(d_i)} - 1)/(log_2(1 + i))$ where $r(d_i)$ is the relevance score of document d_i at position i in the ranking and N_i is a normalization factor. Using NDCG allowed us to precisely measure relevance in cases of tied top $Score(q,d)$ values, giving multiple documents at rank position one. Table 1 shows precision and recall (scaled from 0 to 100) measured on our test queries. We breakdown the queries based on $maxc$, the maximum number of clicks any document received for each query. 87% of queries have $maxc$=1 and this is worst performing segment with a difference of 8.9 in precision compared to the segment with $maxc$=2. Our task here is to study noise reduction methods for queries with $maxc$=1.

Table 1. Accuracy, Recall and % Queries for different $maxc$.

Query segments on $maxc$	1	2	3-5	6-10	> 10
Precision (Avg. $NDCG@1$)	42.5	51.4	54.3	59.9	68.6
Recall (of 12K judged queries)	7.9	3.6	5.8	3.9	26.8
% Queries (out of 75 million)	87%	7%	4%	1%	1%

3. NOISE REDUCTION FEATURES

We now introduce a set of click data features for noise reduction.

3.1 Query Entropy (QE)

Query entropy captures the randomness of clicks associated with the query on different documents and has been explored in [5] to measure variability in result clicks for a query. Formally, we define query entropy similarly to information theory [9] computed over its normalized clicks as: $e(q) = 2^{-\sum_d nc \cdot log(nc)}$ where nc is click count normalized for a query q so as to range from 0 to 1.

3.2 User Engagement

Trail Length (TL): Search trails, such as those in [3], include navigation after the clicked result. The presence of trails shows that the clicked document was not abandoned. For each query-document, we compute average trail length across all of its clicks.

Dwell Time (DW): The time spent by a user on the clicked result has been used as a click feature in previous work [1,3]. For each query-document, we compute average dwell time for all clicks.

Session Utility (SU): We also studied the session level utility of each click. We define a useful session as one ending with a dwell time exceeding 30 minutes following a result click (or click plus trail traversal). Inspecting the data suggests that this is indicative of users finding a useful page and stopping the search. For each query-document, we counted clicks leading to a session end.

3.3 Query Matching Using Title and URL

The title score is % $query\ terms\ in\ Title$ (QTS) and URL score is % $query\ terms\ in\ URL$ (QUS). Scores are computed for each query-document following normalization of the query, title and URL by lowercasing, removing punctuation, etc.

4. EXPERIMENTAL RESULTS

We now present the results of sampling click data for the $maxc$=1 segment using features from Section 3. We measure precision (P) in terms of $NDCG@1$, recall (R) in terms of the percentage of test queries sampled and the sampled data size (S%) relative to that of $maxc$=1 (i.e., relative to 147 million query-document pairs). To combine P and R, we compute the F-measure (F_β), and set β to 0.33 and 0.25 to reflect the preference of users for precision over recall (see [2,7]). We study the effect of features and feature combinations and report results in Table 2. Significant differences in precision (at $p < .01$) using independent measures t-tests are shown. Significant improvements and decrements over all queries with $maxc$=1 appear under P in bold and italics respectively.

Results show that query entropy is effective (see P and F) and that by combining features we can obtain 22 million query-document pairs (15% of the 147M pairs in the segment) maintaining a recall of 43% and precision of 51.2 (up 20% from 42.5). We obtain high-quality clicks from $maxc$=1 segment that perform similarly to clicks with more evidence from the $maxc$=2 segment (Table 1).

5. CONCLUSIONS AND FUTURE WORK

We have presented a study of different features for determining high-quality clicks in potentially noisy click data. When features

Table 2. Sampling click data using various features.

Feature	P	R	S%	$F_{0.33}$	$F_{0.25}$
All queries ($maxc$=1)	42.5	100	100 %	45.1	44.0
Each feature separately					
Query Entropy > 2	*36.1 (-6.4)*	29.5	58 %	35.4	35.7
Query Entropy (1,2]	37.3 (-5.2)	22	20 %	34.9	35.8
Query Entropy = 1	**48.7 (+6.2)**	48.5	22 %	48.6	48.7
Trail Length = 0	*32.7 (-9.8)*	94.3	73 %	35.0	34.0
Trail Length > 0	42.5 (+0)	52.4	27 %	43.3	43.0
Dwell Time ≤ 30s	*34.8 (-7.7)*	59	42 %	36.3	35.7
Dwell Time > 30s	41.5 (-1)	77.4	58 %	43.5	42.7
Session Utility = 0	40.9 (-1.6)	96.5	95 %	43.4	42.3
Session Utility > 0	43.3 (+0.8)	13.1	5 %	35.1	38.1
Url Score = 0	*27.6 (-14.9)*	69.9	51 %	29.4	28.6
Url Score ≥ 50	43.9 (+1.4)	67.2	40 %	45.5	44.8
Title Score = 0	*19.9 (-22.6)*	59.6	36 %	21.3	20.7
Title Score ≥ 50	42.9 (+0.4)	82.6	59 %	45.1	44.2
Feature combinations in addition to Query Entropy = 1					
Trail Length > 0	**52.6 (+10.1)**	19.9	7 %	45.2	48.0
Session Utility > 0	**59.8 (+17.3)**	4.8	2 %	28.5	36.4
URL Score ≥ 50	**54.3 (+11.8)**	25.6	9 %	49.0	51.0
Title Score ≥ 50	**50.7 (+8.2)**	34.3	12 %	48.3	49.3
((QTS or QUS) ≥ 50) and ((SU or TL) > 0)	**55.1 (+12.6)**	17	6 %	45.0	48.7
((QTS or QUS) ≥ 50) or ((SU or TL) > 0)	**51.2 (+8.7)**	42.7	15 %	50.2	50.6

are used independently, only query entropy provides good results. When other features are combined with query entropy, we obtained a further increase in precision of 20% or more. In future work, we will study the impact of these features for all queries, specifically trying to improve the accuracy of other low volume segments (e.g., $maxc$=2). We will also study ways to incorporate these and similar features into Web search ranking algorithms.

6. REFERENCES

[1] Agichtein, E., Brill, E. & Dumais, S. Improving web search ranking by incorporating user behavior information. In *SIGIR*, 19-26, 2006.

[2] Al-Maskari, A., Sanderson, M. & Clough, P. The relationship between IR effectiveness measures and user satisfaction. In *SIGIR*, 773-774, 2007.

[3] Bilenko M. & White, R.W. Mining the search trails of surfing crowds: identifying relevant websites from user activity. In *WWW*, 51-60, 2008.

[4] Deng H., King I. & Lyu, R.M. Entropy-biased models for query representation on the click graph. In *SIGIR*, 2009.

[5] Dou, Z., Song, R. & Wen, J.R. A large-scale evaluation and analysis of personalized search strategies. In *WWW*, 2007.

[6] Gao J. et al. Smoothing clickthrough data for web search ranking. In *SIGIR*, 355-362, 2009.

[7] Järvelin, K. & Kekäläinen, J. IR evaluation methods for retrieving highly relevant documents. In *SIGIR*, 2000.

[8] Joachims, T. Optimizing search engines using clickthrough data. In *SIGKDD*, 133-142, 2002.

[9] Shannon, C.E. Prediction and entropy of printed english. *The Bell System Technical Journal*, 30: 50-64, 1950.

[10] Xue, G. et al. Optimizing web search using web click-through data. In *CIKM*, 118-126, 2004.