Through-the-Looking Glass: Utilizing Rich Post-Search Trail Statistics for Web Search

Alexey Tolstikov, Mikhail Shakhray, Gleb Gusev, Pavel Serdyukov
Yandex
16 Leo Tolstoy St., Moscow, 119021 Russia
{atolstikov, smikler, gleb57, pavser}@yandex-team.ru

ABSTRACT

With increasing popularity of browser toolbars, the challenge of employing user behavior data stored in their logs rises in its importance. The analysis of post-click search trails was shown to provide important knowledge about user experience, helpful for improving existing search systems. However, the utility of different trail properties for improving existing ranking models is still underexplored. We conduct a large-scale study and evaluation of a rich set of search trail features in realistic settings and conclude that a deeper investigation of a users experience far beyond her click on the result page has the potential to improve the existing ranking models.

Categories and Subject Descriptions: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

Keywords: user behavior features; search trails; dwell time.

1. INTRODUCTION

In recent years, user behavior data plays an increasingly important role in different IR tasks. The most well-known way to derive evidences of users' preferences and satisfaction is analyzing web search click logs. Though they provide the vast amount of implicit user feedback, its value and reliability is markedly limited, since a large part of users' activity takes place beyond their clicks on search results.

With increasing popularity of browser toolbars it becomes possible to partially compensate for the lack of post-click behavior data by toolbar logs that store browsing actions of their users. It was shown [1] that basic statistics of user interactions with web pages, such as dwell time, may serve as high-performing features for a document ranking model. However, the entire sequence of pages visited by a user with the same information need after she made a click on a result page, a so-called "post-query search trail", is not well-studied as a source of features with a potential to improve the ranking of documents participating in the trails. We assume that a substantial analysis of search trails may help to further improve existing retrieval models in comparison to already well-known features, such as dwell-time.

In the current paper we provide a large-scale investigation of different properties of search trails that continues previous studies on user behavior data and its utility for web search. Following [7], we represent search trails as tree-like structures with the clicked results as their roots and chains of forward hyperlink transitions as their branches. Being a tree, a search trail possesses its characteristics: nodes count, depth, breadth, average branch length. In addition to these trail features, we also study and evaluate some new ones, including the number of trail steps with a known inactivity duration observed after them. Some of these properties of browsing trails were investigated previously in theoretical studies such as [7], but, to the best of our knowledge, their utility for web retrieval has not been yet evaluated by IR metrics. Being aggregated at the document or domain level of the clicked result, most of features significantly improve the performance of a baseline retrieval model that utilizes state-of-the-art post-search trail features. This result supports our above-mentioned assumption that by going deeper beyond dwell time, we are able to learn more about clicked results' relevance.

To sum up, the contributions of the paper are: (1) we conduct a large-scale study of a broad family of search trails' features and their utility in web search, (2) we reveal that a substantial study of search trail characteristics may provide some additional evidences essential for information retrieval tasks.

2. RELATED WORK

From a search engine perspective, the most practical way to incorporate user behavior data into an existing ranking system is likely to be the development of new features reflecting different qualities of user interactions with a website. One of the first papers on exploiting user behavior features extracted from browsing logs for improving quality of a competitive ranking is [1]. Among other behavior features, the authors investigated some basic statistics of user interactions with web pages including different variations of dwell time. More subtle evidences of user browsing experience may be obtained by the analysis of scrolling and cursor movements [4]. As in this paper, we also look "beyond dwell time" while deriving evidences of user experience, but we also go far beyond the first page in the trail. Another possible approach to utilize user behavior data is developing a text retrieval scoring based on language models of initial

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2507890.

queries that lead to the examined document via search trails mined from a toolbar log [2]. The cumulative value of search trails was shown to exceed the value of their origin and destination pages when comparing them by different measures such as relevance, topic coverage, topic diversity, novelty and utility [8]. In our study, we represent trails as tree-like structures as proposed in [7]. We also adopt a part of basic graph properties considered in that study. Some of these properties proved their utility in the task of best trail finding [5]. The binary feature of the clicked result that indicates the presence of any post-click trail was used for training a classifier for detection of noisy clicks [3].

3. DATA

All the experiments reported in this paper were performed by utilizing the user behavior data stored in the anonymized log of a popular search engine browser toolbar used by millions of people across different countries. Each record in this log contains the (anonymous) toolbar user identifier, a timestamp, and the details of the browsing action, such as a query submitted by the user, URL of a visited page, or closing the browser window. We extracted all the records stored in the toolbar log during the three-month period from 11 December 2012 till 10 March 2013. This data contains 3,0B user queries, 5,3B search trails, and 16B page visits covering 2,7B different documents.

From the obtained data, we extracted *search trails* that start with a user query and consist of the sequential web page visits by the same user likely to be related to the same information need. To reduce the noise coming from pages unrelated to the user information need expressed by the initial query, we terminated a search trail in the case of one of the following events: (1) user submitted a new query, (2) user navigated home page, entered URL into the browser address bar, or transitioned to a web page by using the browser bookmark, (3) there were no browsing activity more than 30 minutes (inactivity timeout), (4) user closed the browser window. This is the list of rules similar to those that define a search trail according to [7] with the exception of the rule "check email or logon to service" that seems to be counter-intuitive, as, in fact, a user may still continue the search task by clicking a hyperlink taking her to a website requiring authentication.

4. SEARCH TRAIL FEATURES

In this section, we briefly describe the way of search trails construction similar to those proposed in [7]. As we already mentioned above, we treat each search trail as a tree-like structure. Nodes of these trees represent unique pages, and directed edges represent user transitions through hyperlinks between them. In such a way, forward user hyperlink transitions are reflected as moves along a tree branch. Besides, if some user repeatedly visits some page previously visited at some preceding trail step, this move is represented as the user transition back to the corresponding tree node previously visited by the user. After that, new pages visited by means of further hyperlink transitions, if any, constitute a new branch of the tree. If the user returns to the result page and clicks on a new document, we initiate a new tree. See an example of a resulting tree-like structure on Figure 1. In the next subsection, we describe the properties that a trail can be characterized by and which we utilize as ranking features later in this paper.

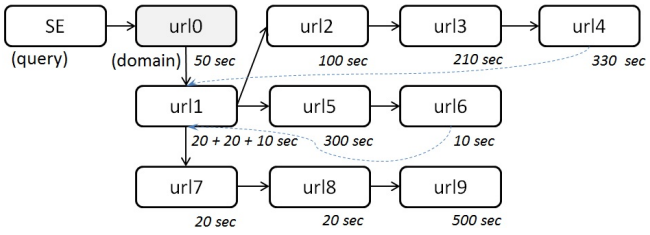

Figure 1: A search trail represented as a tree. Nodes = 10, depth = 4, breadth = 3, branch length = 3, steps = 12, revisits = 2, time = 1590, satisfied steps = 6, long steps = 3.

4.1 Graph Features

- *Nodes* count. The total number of tree nodes corresponds to the number of unique pages visited by the user during the post-click search trail. Large values of this feature may indicate that the first trail page served by the search engine among its results was not enough to satisfy the user information need forcing her to seek deeply by following link transitions. On the other hand, large values of this feature are more typical for trails initiated by informational queries, whose information need can not be fully covered by any single web page.

- *Depth* is the distance between the tree root and the most distant node, where the *distance* between two nodes of a tree is the number of edges in the shortest path connecting these two nodes by edges of the tree. Deep trees are presumably more typical for trails that represent browsing on a website, whose pages are served in chains sequentially formed by forward and backward hyperlinks. It might be the case for the information that is designed to look it through by traversing an ordered list of web pages.

- *Breadth* of a trail tree is the the number of its leaves. *Leaves* represent destination pages whose visits were never followed by a forward hyperlink transition. Trail breadth coincides with the number of branches, the latter quality was considered in [7]. Large values of this feature may indicate that the underlying information need has many aspects, the user perform seeking in an exploratory manner, or the domain containing the trail pages is designed in an inconvenient way.

- Average *branch length*. We split a search trail into segments, each next segment starts with a repeated visit of a previously visited page and constitutes a chain of sequential forward hyperlink transitions. For each chain, we find its *length*, which is the number of edges forming that branch the chain is made of. We ignore chains of length 1 since they do not initiate new tree branches. Average branch length is obtained as lengths averaged over all chains, which correspond to different tree branches. What is worth mentioning, this value is also equal to ((nodes-1)/breadth)+1.

nodes		depth		breadth		branch length		steps	
Private life	2.20	Private life	1.91	Sports	1.26	Private life	2.65	Private life	2.68
Sports	2.15	Rest	1.85	Rest	1.24	Rest	2.63	Sports	2.64
Rest	2.14	Sports	1.83	Private life	1.24	Automobiles	2.59	Rest	2.63
Automobiles	2.06	Automobiles	1.80	Automobiles	1.21	Business	2.58	Automobiles	2.54
Business	1.98	Business	1.76	Employment	1.21	Sports	2.58	Business	2.42

revisits		diversity		sat. steps		long steps	
Entertainment	0.70	Computers	1.13	Society	0.24	Rest	1.15
Sports	0.56	Entertainment	1.12	Media	0.20	Society	1.11
Private life	0.55	Employment	1.12	Entertainment	0.20	Automobiles	1.10
Rest	0.54	General	1.12	Science	0.20	Sports	1.08
Automobiles	0.53	Culture	1.09	Rest	0.19	Employment	1.05

Table 1: Topics with the highest mean value of each of the trail-based features aggregated by domains.

4.2 Movement features

Besides the above features that represent properties of the trail tree itself and thus depend only on its topology, there are also some other trail features reflecting different qualities of the user walk on the trail tree.

- Number of *steps* of a trail is the total number of transitions made by the user while browsing along the trail. This feature is similar to nodes count introduced above, but differs from it in that we also count all the repeated page visits to compute this number.
- *Revisits* count is the number of repeated page visits made by the user along the trail. Revisits count may be considered as a measure of trail intricacy. In fact, large values of revisits count signify that the user often returned to previously visited pages, either to navigate some new pages linked from there, or to learn some information that she were not able to learn at the first visit of these pages.
- *Diversity* is the number of different second-level domains represented by trail pages.
- *Satisfied steps* count and *long steps* count are the numbers of trail steps followed by 30 and 300 seconds of inactivity respectively. Thereby, we define satisfied trail steps in a similar way as satisfied clicks are usually defined (see, ex., [6]). Satisfied steps correspond to pages that turn out to be worth of a remarkable amount of user attention.

Figure 1 captures an example of a search trail and reports values of all its features described above.

4.3 Feature aggregation

After all the trail features were extracted for each individual trail, we aggregated them over all the trails via one of the two possible ways: at the level of the first document of the search trail (*URL-level aggregation*) and at the level of that document domain (*domain-level aggregation*). As a result of each aggregation type, we obtained samples of search trails associated to either document, or domain. For each property of a trail described above, we calculated its mean (*av*), standard deviation (*std*), 10th and 90th percentile (*10th*, *90th*), maximal and minimal values (*min*, *max*) and used them as features in our ranking model. In the next section we conduct a study of how the described features depend on the topic of the web page domain.

5. FEATURES AND DOMAIN TOPICS

In this section we study the distribution of search trail features conditioned by different topics of their initial web pages. To this end, we use a proprietary database of domains manually categorized by their topics. We implemented a Naive Bayes classifier trained on this data by employing unigram features of domain pages. This classifier assigns each second-level domain whose documents are represented in our data set of search trails with some topic chosen from among the topics of the categorized data base. For each mean type feature aggregated on a domain level (see Section 4.3), we calculated its mean value over all the pages within the same topic. This way, we attributed to each topic mean values of each of the considered features. For a given feature, we rank all the topics according to that mean value and report the obtained results in Table 1.

As we can see, some topics have natural interpretations of falling into the corresponding top lists when be measured by the trail features. For example, a user who browses a website devoted to car selling cannot really know in advance which particular car she looks for. A user also likely explores different pages devoted to various rest facilities before she learns the possible opportunities. Similar observations can be made for such features as depth, breadth, and steps. The highest number of satisfied steps are gained by such topics as Society, Media and Science, whose content mostly consists of articles served to read them deeply in. Besides the results reported in Table 1, we also revealed some notable regularities at the bottom of topic lists. Among the topics having small values of satisfied steps are Private life and Automobiles whose steps number are rather large. Despite the large number of visits, a user is not likely to stay for a long time at pages of that topics domains. These results indicate that trail features may contain some information on domain topic that may be of use for a search system. In the next section, we describe evaluations of the trail features and their utility for web search.

6. EVALUATION

While evaluating trail features, we relied on a large-scale data set of user queries randomly sampled from the web search of a major search engine. For each query, top documents served by the world's leading search systems were explicitly annotated by professional judges with labels from among "perfect", "excellent", "good", "fair", and "bad". In total, this data set contains 50K queries and 1,5M labeled

query	Basic	Basic+Domain		Basic+URL	
all	57.57%	57.95%	**+0.66%**	58.05%	**+0.82%**
1 words	67.62%	67.78%	+0.23%	67.47%	-0.21%
2 words	64.72%	64.89%	+0.27%	64.91%	+0.29%
3 words	59.08%	59.63%	+0.93%	59.48%	+0.67%
≥ 4	50.15%	50.75%	+1.2%	50.92%	+1.54%
popular	66.91%	67.16%	+0.36%	67.11%	+0.30%
unpop.	49.88%	50.34%	+0.91%	50.55%	+1.33%

Table 2: NDCG@10 scores gained by baseline model, with employing URL-aggregated and domain-aggregated features. Forming 45.18% of the data set, queries of rate ≥ 10 per week are called popular. Differences in bold are statistically significant at the 0.99% confidence level.

query	Basic	Basic+Domain		Basic+URL	
buc. 1	44.11%	44.60%	+1.1%	44.98%	+1.97%
buc. 2	59.61%	60.01%	+0.67%	59.85%	+0.39%
buc. 3	65.84%	66.12%	+0.42%	66.09%	+0.37%
buc. 4	67.01%	67.34%	+0.49%	67.31%	+0.45%

Table 3: NDCG@10 scores gained at four different levels of data availability from bucket 1 (least availability) to bucket 4 (highest availability).

query-documents pairs. In all the evaluations, we trained Friedman's gradient boosting decision trees as a ranking model. We compared the performance of suggested features to the performance of the following baseline feature set (*Basic*): a variant of BM25 score, PageRank, CTRs aggregated at domain and document levels, and 7 modifications of dwell times investigated in [1, Table 4.1]: TimeOnPage — TimeOnDomain and AverageDwellTime — DomainDeviation. This baseline is thus strong enough, easily interpretable, and includes a wide range of currently known dwell-time based features.

We divided all the queries of the data set into two equal parts, first one for learning models and the second one for evaluation. In Table 2, we report performance of the three models trained by using: (1) Basic set of features; (2) Basic and domain-aggregated trail features, and (3) Basic and URL-aggregated trail features. Both domain- and URL-aggregated features demonstrate their benefit on the test collection. A model trained on Basic features without 7 modifications of dwell times performs at the level $NDCG@5 = 55.9\%$. Hence, URl-based trail features earn 0.82% of quality additionally to $2,9\%$ gained by dwell times. We also measured the performance of the three models on different query classes separately. We revealed that search trail features contributed even more for long and rare queries. We explain it as follows: being aggregated by documents and domains, our search trail features propagate important evidences of user experience to more difficult cases where baseline user behavior features are sparse and thus not informative. In order to confirm this guess, we split all the queries in the test into the four nearly equal parts representing different level of availability of search trail data measured in the number of trails of at least 2 steps that were initiated by the given query. The obtained results reported in Table 3 indicate that trail features gain even more for queries with a lack of search trails. In Table 4, we report top 10 features according to their contribution, which is measured

#	Basic+Domain		Basic+URL	
1	QueryDomCTR	20.2	QueryDomCTR	21.7
2	BM25	17.7	BM25	20.2
3	QueryUrlCTR	14.2	QueryUrlCTR	13.1
4	QDwellTimeDev	11.0	QDwellTimeDev	10.9
5	PageRank	5.2	*AvSatSteps*	5.1
6	*AvSatSteps*	2.5	PageRank	4.7
7	AvDwellTime	2.2	TimeOnDomain	2.6
8	DwellTimeDev	1.7	CumulativeDev	1.8
9	*90thDwellTime*	1.4	*90thDwellTime*	1.7
10	*10thDwellTime*	1.3	AvDwellTime	1.7

Table 4: Top 10 features according to their contribution.

in weighted improvement of loss function over all employments of a feature during the learning process. Search trail features are marked out by *italic*.

7. CONCLUSION

We performed a large-scale study of post-click search trails and their utility in web search. We considered a rich set of trail properties as a potential source of information about user experience taking place far beyond her click on the initial result page. A detailed evaluation demonstrates remarkable contribution of search trail features to a strong baseline retrieval model. To the best of our knowledge, most search trail features were not previously evaluated by IR metrics. We believe that future substantial analysis of search trails including investigation of new trail qualities and different ways of their aggregation may help to even further improve existing retrieval models in comparison to already known user behavior features.

8. REFERENCES

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

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

[3] Q. Guo and E. Agichtein. Smoothing clickthrough data for web search ranking. In *SIGIR*, pages 355–362, 2009.

[4] Q. Guo and E. Agichtein. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *WWW*, pages 569–578, 2012.

[5] A. Singla, R. White, and J. Huang. Studying trailfinding algorithms for enhanced web search. In *SIGIR*, pages 443–450, 2010.

[6] K. Wang, T. Walker, and Z. Zheng. Pskip: estimating relevance ranking quality from web search clickthrough data. In *KDD*, pages 1355–1364, 2009.

[7] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *WWW*, pages 21–30, 2007.

[8] R. W. White and J. Huang. Assessing the scenic route: measuring the value of search trails in web logs. In *SIGIR*, pages 587–594, 2010.