Introducing Search Behavior into Browsing Based Models of Page's Importance

Maxim Zhukovskiy Andrei Khropov Gleb Gusev Pavel Serdyukov
Yandex
16 Leo Tolstoy St., Moscow, 119021 Russia
{zhukmax, akhropov, gleb57, pavser}@yandex-team.ru

ABSTRACT

BrowseRank algorithm and its modifications are based on analyzing users' browsing trails. Our paper proposes a new method for computing page importance using a more realistic and effective search-aware model of user browsing behavior than the one used in BrowseRank.

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

General Terms: Algorithms, Experimentation, Performance

Keywords: Page authority, BrowseRank, queries, web search

1. INTRODUCTION

In order to improve the web search quality, many approaches to measuring page importance were proposed: based on link analysis, user behavior, keyword-based features, etc. BrowseRank [1] is a classic link-based authority measure taking user browsing paths into account. According to this method, the importance of a page equals to its weight in the stationary distribution of a continuous-time Markov process on the user browsing graph. Despite undeniable advantages of BrowseRank and its generalizations [2, 3], these algorithms have the following weaknesses.

1) *As defined in Browserank, a browsing session ends if the user submits a new query.* This assumption is illustrated by Fig. 1. However, inside one *search session* the user can submit several queries in order to clarify the first query and continue browsing pages supposedly related to the ones visited after the first query, since all queries and pages visited in between belong to the same information seeking process. That is why we consider both pages and queries to be parts of browsing sessions. Moreover, the pages the user surfs after submitting such "clarifying" queries are often more important for the user than the first visited pages in the browsing session. Therefore, we need to take into account at which stage in the entire information seeking process the page was visited when trying to deduce how important it is from the browsing sessions.

2) *Probability of choosing a page at random with reset probability by the user does not depend on the average position of the page in a browsing session.* Let page p_1 be the *destination page* or the *original page* (see Fig. 1) in a large part of browsing sessions. Let page p_2 be in the middle parts of the same sessions. According to the existing studies [4], we conclude that there is a better chance that page p_1 is more relevant to the corresponding queries than page p_2. But the BrowseRank score is not sensitive to this effect. As we show in Section 3, by tuning the damping factor for a page we can influence its probability in the stationary distribution of the Markov process underlying BrowseRank algorithm.

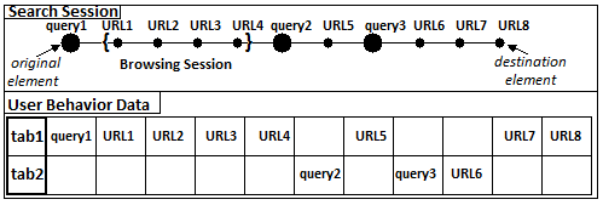

Figure 1: Types of sessions.

In this work we present the algorithm called *Search-aware BrowseRank*. Our algorithm exploits a *modified user browsing graph* with additional vertices which represent queries and additional edges which connect queries and pages. This modification serves to solve the first of the mentioned above problems by considering queries as parts of browsing sessions and distinguishing between queries having different positions inside one browsing session (which happens to be also a search session in this case). The second disadvantage is overcome by making the damping factor depend on the average position of the page in a session. In order to show the effectiveness of Search-aware BrowseRank, we compare it with the classic BrowseRank on a large sample of the user browsing graph.

To sum up, the contributions of this paper are the following. 1) We introduce a new definition of a browsing session and propose a new algorithm of computing page importance based on this definition. Our algorithm observably depends on the user's search behavior within the framework of one browsing session. 2) We test our algorithm on a large data set and demonstrate that our algorithm is better than BrowseRank.

2. SEARCH-AWARE BROWSERANK

Let us consider a user and its browsing history. We add queries in the browsing sessions defined by Liu et al. [1] to make them user's search behavior dependent. In contrast to Liu et. al., who considered that user's session ends with submitting a new query, we represent these sessions as sets of pages and queries with the following properties. The session starts with a web page or with a query. Let $p_{1,1}, ..., p_{1,j_1}, q_1, p_{2,1}, ..., p_{2,j_2}, q_2, ...$ be the first elements of

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

the session (there are both pages and queries in this set). Let the last considered record in the user's log to be made at time t (either about visiting of page $p_{k,i}$, or about submitting query q_k), and the next record to appear at time $\tilde{t}$.

Let the difference $\tilde{t} - t$ be not greater than 30 minutes. If the record made at time $\tilde{t}$ describes either the transition from one of the pages p from $\{p_{1,1}, p_{1,2}, ..., p_{k,i}\}$ to page $p_{k,i+1}$ by clicking a hyperlink, or clicking the page $p_{k,i+1}$ after submitting one of the queries q from $\{q_1, q_2, ..., q_k\}$, then we add $p_{k,i+1}$ into the session and call either $\{p, p_{k,i+1}\}$ or $\{q, p_{k,i+1}\}$ the pair of neighboring elements. If the record at time $\tilde{t}$ describes submitting query q_{k+1} then we add this query in the session and call $\{p_{k,i}, q_{k+1}\}$ the pair of neighboring elements (if the record made at time t describes submitting q_k then $\{q_k, q_{k+1}\}$ is the pair of neighboring elements).

In all the other cases page $p_{k,j_k} := p_{k,i}$ (or query q_k) is the *destination* page (or query) of the session. New browsing session starts from the first element (either from a query or from a page) in the record made at time $\tilde{t}$.

For each page p (or query q) from the user's logs, we denote the number of sessions this page (or query) starts with by $s(p)$ (or $s(q)$). For each pair of neighboring elements $\{v_i, v_{i+1}\}$ from a session, we denote the number of sessions containing such a pair of neighboring elements by $I(v_i, v_{i+1})$. We denote elements of a browsing session S by $v_1(S), ..., v_{k(S)}(S)$. Corresponding time moments of records from the log we denote by $t_1(S), ..., t_{k(S)}(S)$ respectively.

We define *Search-aware browsing graph* $G = (V, E)$ as follows. The set of vertices V consists of all the web pages V_{page} and all the queries V_{query} mentioned in the log and contains one additional vertex x. The set of directed edges $v_i \to v_{i+1}$ in E represents all the ordered pairs of neighboring elements $\{v_i, v_{i+1}\}$ in the browsing sessions. Set E contains additional edges from the last pages of the sessions to x. We set $s(x) = 0$. Let $I(v, x)$ be equal to the number of visits of v if $v \to x \in E$. When modelling user's behavior on Search-aware browsing graph we consider *reset probability* (probability of starting a new browsing session) and *transition probability* (probability of clicking a hyperlink or submitting a query inside the current session). The reset probability $\sigma(v)$ for $v \in V$ equals $s(v)/\left(\sum_{\tilde{v} \in V} s(\tilde{v})\right)$. Let $v_i \to v_{i+1}$ be in E. The transition probability $\omega(v_i \to v_{i+1})$ equals $I(v_i, v_{i+1})/(\sum_{v \in V:\, v_i \to v \in E} I(v_i, v))$.

In order to define the expected staying time of the user on a page p, we consider the set $T(p)$ of observable staying times $t_{j+1}(S) - t_j(S)$ such that $p_j(S) = p$. We define the estimated staying time $Q(p)$ of page p by the sample $T(p)$ in the same way as in [1]. In other words, we propose that the distribution of the staying time is the sum of a Chi-square distribution and an exponential distribution.

Search-aware BrowseRank algorithm is defined in a similar manner as in [1]. Its outcome is the stationary distribution of the continuous-time Markov process on the Search-aware browsing graph. Denote $\alpha(p)$ ($\alpha(q)$) the probability of moving by a random edge by the user being at the page p (or query q). Let $v \in V$. Let $s_*(v)$ be the fraction of sessions containing v as the first element, $s^*(v)$ be the fraction of sessions containing p as the final element. Let $a, b, c \in [0, 1]$ be the tunable parameters. We set $\alpha(v)$ to be the linear function $as_*(v) + b(1 - s_*(v) - s^*(v)) + cs^*(v),\ v \neq x$.

Search-aware BrowseRank for $v \in V$ equals $Q(v)\pi(v)$, where $\pi(v)$ is the solution of the following system of linear equations:

$\pi(v) = (1 - \tilde{\alpha}(v))\sigma(v) + \alpha(v) \sum_{\tilde{v} \neq x:\, \tilde{v} \to v \in E} \omega(\tilde{v} \to v)\pi(\tilde{v})$, where $\tilde{\alpha}(v) = \alpha(v)(1 - \pi(x))$ (these equations hold for $v = x$ as well). This definition differs from the definition of BrowseRank only in the choice of graph and damping factor $\alpha(v)$. The systems of linear equations defining $\pi(v)$ for BrowseRank and Search-aware BrowseRank are the same.

3. RESULTS AND CONCLUSIONS

All experiments are performed with users' behavior data recorded in December 2012 in bowser toolbar log of the popular search engine Yandex[1]. There are ≈800M pages and ≈4.41B transitions in the log. For ranking evaluation we sampled 1000 queries from the queries submitted by real users. For each query, a set of URLs was evaluated by professional assessors hired by the search engine. The relevance scores were selected from among the editorial labels: *Perfect, Excellent, Good, Fair, Bad*. We compare Search-aware BrowseRank (SBR) with classic BrowseRank in the following way. These algorithms are combined linearly by ranks with BM25. In order to compensate the difference in scales of Browserank and BM25 scores, the parameter of linear combination is chosen in such a way that average values of the feature and BM25 multiplied by the parameter are equal. Let us introduce the ranking performance on metrics NDCG@10 over the algorithms with different sets of parameters:

$(0.85, 0.85, 0.85)$	$(0.7, 0.85, 0.85)$	$(0.85, 0.85, 0.7)$
0.79432	0.77251	0.76325

$(0.85, 0.7, 0.7)$	$(0.7, 0.7, 0.85)$	$(0.7, 0.8, 0.9)$	$(0.9, 0.8, 0.7)$
0.75987	0.80072	0.8092	0.7891

Table 1: Performance of SBR with parameters (a, b, c).

For BrowseRank, we get NDCG@10≈0.7561. Our experimental results show how tuning the damping factor for pages can influence the ranking performance of the algorithm. The best result is achieved for the case $a < b < c$. Therefore, we believe that tuning the damping factor as a function of a position of a page in sessions and increasing the weight of destination pages in this function can increase the ranking performance of the algorithm.

Our results may be used by commercial web search engines for improving their search quality. It is interesting to continue studies of the damping factor in the above mentioned settings.

4. REFERENCES

[1] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, H. Li, *BrowseRank: Letting Web Users Vote for Page Importance*. Proc. SIGIR'08, pp. 451–458 , 2008.

[2] Y. Liu. T.-Y. Liu, B. Gao, Z. Ma, H. Li, *A framework to compute page importance based on user behaviors*, Inf Retrieval, 13: 22–45, 2010.

[3] B. Gao, T.-Y. Liu, Z. Ma, T. Wang, H. Li, *A General Markov Framework for Page Importance Computation*, Proc. CIKM'09, pp. 1835–1838, 2009.

[4] R. W. White, J. Huang. *Assessing the Scenic Route: Measuring the Value of Search Trails in Web Logs*, Proc. SIGIR'10, pp. 587–594, 2010.

[1] yandex.com