Search Logs Mining: Survey

Vivek Bhojawala and Pinal Patel

Abstract Search engine process millions of query and collect data of user interaction every day. These huge amount of data contains valuable information through which web search engine can be optimized. Search engine mostly relies on explicit judgement received from domain experts. To survive the competition search engine must understand user's information needs very well. Search logs provide implicit data about user's interaction with search engine. Search logs are noisy, they contain data of both successful search and unsuccessful search. The challenge is to accurately interpret user's feedback to search engine and learning the user access patterns, such that search engine will better be able to cater the user's information needs. User feedback can be used to re-rank the search result, query suggestion and URL recommendation.

Keywords Search logs · Click-through · Implicit feedback · Search engine · Web search ranking · Search destination · Entropy · Information goal · User behavior · Search history · Learning from user behavior

1 Introduction

Web search engine mostly consists of explicit judgement received by domain expert. Web is dynamic, new websites are created every day and also new queries with different information need arrives. To prepare such relevance judgement and maintain it up to date is an expensive and time consuming task. Search logs contains implicit user feedback which can be used as relevance judgement. To use implicit feedback as relevance judgement we need to understand how user interacts with search engine. From search log we can: (1) Re-rank search result: search engine uses

V. Bhojawala (✉) · P. Patel
Government Engineering College Gandhinagar, Gandhinagar, India
e-mail: vivek141@gecg28.ac.in

P. Patel
e-mail: pinalpatel@gecg28.ac.in

S.C. Satapathy and S. Das (eds.), *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2*,
Smart Innovation, Systems and Technologies 51, DOI 10.1007/978-3-319-30927-9_4

implicit feedback of user to rank the result. (2) Find Ambiguous query: User queries are often ambiguous. Single query contains multiple meanings. (3) Positional bias of results: How User interacts with top 10 results returned by a Search engine. (4) Interpreting User behavior: Differentiating user's behavior for successful and unsuccessful search. (5) modeling user's implicit feedback: it can be modeled at query level, session level or task level. (6) Dynamic search result: based on past search of user search result contains both old and new results based on relevancy. (7) Measure Efficiency of search engine: Search engine is producing relevant result to query. (8) Query suggestion: based on implicit feedback similar queries are suggested to user for (9) URL suggestion: based on URL visited by other user for same query. (10) Prediction of user action: based on most recent interaction search engine can make prediction of upcoming information need of user.

2 Web Search and Search Logs

This section contains: (1) comparison of classic information retrieval system and classical information retrieval augmented for web search. (2) Classification of query. (3) web search behavior of user. (4) Example of anonymized search log. (5) Measuring entropy of search logs.

2.1 Web Search Fundamental [1]

In information retrieval system user is having specific information needs [1]. That information need is converted into query and submitted. Submitted query is matched against collection of documents (corpus) with certain rules and most relevant documents are returned back to the user.

Web search contains a little bit different structure from information retrieval system. In web search main difference is users perform tasks rather than specific information search, each task requires some information and that information need is converted into verbal form and then it is submitted as a query to search engine. Now search engine will return results based on some rules and relevance of the documents to the query submitted by the user. User examines the results and now user will compare relevance of URL returned by search engine to the information need. If result matches with information need then user is satisfied and stops searching for that topic otherwise user will reformulate the query and repeats the process until user gets required information need. After a fair amount of query reformulation user may abandon the search task.

Web queries can be: (1) Navigational: User is searching for particular web site URL. Like Login page of Gmail, official website of android Marshmallow, home page of apple iPhone etc. Navigational query is generally satisfied with single URL click. (2) Information: User search for information which is based on facts that can be present on multiple websites. Like height of the Mount Everest, capital of India, planned cites of India etc. Information queries are generally satisfied by multiple URL clicks. (3) Transaction: Transactional query involves searching for URL such that user can perform more action on that website. Like 'buy nexus 5' will return multiple e-commerce websites which sells nexus 5, user selects particular website and performs other interaction with website like applying coupon, giving address and contact information for shipping of nexus 5 and at last user selects one of the payment option available on the website. Payment option may redirect user to a particular bank website for online payment.

3 Understanding the Search Logs

This section contains: (1) Interpreting click through data. (2) Measuring retrieval quality of search engine. (3) Relationship between searcher's query and information goal (4) Features representing user behavior.

3.1 Interpreting Click Through Data [2]

Using search logs as implicit feedback is difficult to interpret correctly and it includes noisy data. To evaluate reliability of click through data, study was conducted to analyze how users interact with Search engine result page and compare their implicit feedback with explicit feedback. To analyze following experiments were performed (1) user views result page from top to bottom? How many abstract do they read before clicking? (2) How implicit feedback matches with explicit feedback constructed by domain experts. For experiment users were asked questions which contained both informational and navigational query. Eye tracker was used to capture eye fixations which is defined as concentration of eye at particular part of web page for 200–300 ms approximately. Fixation measures interestingness of URL.

Results of experiment shows: (1) user views first two links of result page equally but number of clicks for 1st links are very high compared to second link. Same behavior observed for 6th and 7th link. (2) User scans results from top to bottom. (3) User does not observe abstracts of all links but more likely to observe abstract of clicked link and link above and below the clicked link. To better understand user behavior on first two links second experiment was carried out. In second experiment, each user was assigned one of the following three conditions: (1) Normal: User was given results directly received from Google search engine. (2) Reversed: Results returned by user were reversed. (3) Swapped: First two results returned by search engine were swapped. Result of experiment shows: (1) in reversed list of result user viewed more abstracts compared to normal shows that order in which relevant results presented does not matter, user views abstracts and clicks according to query. (2) When top two results were swapped even if 2nd result was more relevant to query, most user clicked 1st link showing trust bias. Trust bias is user's trust on particular search engine.

Important deductions from experiments. (1) If user clicks on particular URL it means that URL is examined by user and it is relevant to query issued by the user. (2) If URL is clicked and URL above that is skipped means that user examined the URL and it is not relevant to query. (3) The rank at which user clicks on the link is also important and it shows relevance to particular user.

3.2 Measuring Retrieval Quality of Search Engine Using Search Logs [3]

To measure retrieval quality of search engine search logs can be directly provided as input to feedback system. User's satisfaction is ultimate goal of search engine, so in order to better serve users search engine needs to measure its own retrieval quality form click through data. Following are the absolute metrics to measure retrieval quality. (1) Abandon rate: It is measured in number of times user issued a query and didn't clicked on any results. (2) Reformulation rate: it is measured in part of query used by successive query in the same session. (3) Clicks per query: Mean no of results clicked for each query. (4) Time to first click: Mean time between queries submitted by user and first click of results. (5) Time to last click: Mean time between queries submitted by user and last click of results.

3.3 Relationship Between Search's Query and Information Goal [4]

For rare and complex information goal user behavior changes significantly, click through rate decreases and query reformulation increases. User query can be specific or general to information needs, success of search depends on search engine's ability to interpret the information needs. Efficiency of search engine can be measured in session length. When search engine is unable to produce relevant result to user query session length increases. Search session contains sequence of queries in chronological time. Session ends when 30 min of inactivity. A detailed observation in search logs leads to observation that user issues more than one queries which are interrelated for single information goal.

Example shows that at time t0 user issued query to buy Samsung mobile online and then user clicks on Samsung mobile home page and from there user picked one particular model, click at time t2 is URL Click indicating that visited URL is not from SERP. After finding particular model of mobile user searched for review of that mobile. After reading review, at t4 user decided to buy a phone at t5 and after seeing result of query user reformulate the query to buy phone from particular e-commerce website at time t6 and finally bought the phone. Click at time t8 is checkout click to fill out payment and shipping details from particular e-commerce website. So we can conclude that user executed all this queries. Search engine returns results according to user's query and query reflects information goal. It is difficult to know precisely user's satisfaction. User's satisfaction can be measured by examining URL visited by user at the end of the session. In this session example, we can say that user is satisfied by examining last URL click which is of checkout which proves that user has bought the mobile which was user's ultimate goal. Other concern about user behavior is parallel loading URLs in browser tabs without reading the content of URL. To identify sessions of parallel tabs, dwell time is observed. If dwell time is below threshold for more than one clicks then user may have this scenario.

To study post query behavior of user for rare information goal two weeks of search engine data was collected. Tail query is defined as queries and URLs observed in second week that were not observed in first week for the information goal initiated in first week, all other queries are non-tail. Results in comparison of tail and non-tail query shows that: (1) Query reformulation rate for tail query is higher than non-tail query because tail query represents rare and specific information need. Length of reformulation represents search engine's ability to understand rare information need of user. (2) When user reformulate the query to be more specific then query length increased as compared to initial submitted query.

3.4 Features Representing User Behavior [5]

Features representing user behavior can be (1) Query Text Features, (2) Browsing Features, (3) Click Through features. Query feature includes: (1) query length: Numbers of words in query. (2) Next query overlap: number of words common with next query. (3) Domain overlap: words common with query and domain. (4) URL overlap: words common with query and URL. (5) Abstract overlap: words common with query and abstract. Browsing feature includes: (1) Dwell time: Time spent on URL. (2) Average dwell time: Average time on the page for single query. (3) Dwell time deviation: Deviation from overall average dwell time on page. Click Through feature includes: (1) Position: Rank at which URL clicked. (2) Click frequency: Number of clicks for this query, URL pair. (3) Click relative frequency: Relative frequency of click for this query and URL. (4) Click Deviation: deviation from observed number of clicks.

4 How People Recall, Recognize, and Reuse Search Results [6]

When a user issues the query, user has certain expectations about how search results will be returned. These expectations can be based on information goal, and also based on knowledge about working of search engine. Such as where relevant results are expected to be ranked based on previous searches of individual user for a specific topic. When a result list of URL is changed according to particular modeling schema, users have trouble re-using the previously viewed content in the list. However study has shown that new relevant result for same query can be presented where old results have been forgotten, making both old and new content easy to find.

4.1 Recall

Two main factors affecting how likely result was to remember (1) position at which result was ranked. (2) Whether or not the result was clicked. Results that were clicked were significantly more likely to be recalled. 40 % of clicked result were remembered, compared with 8 % of results that were not clicked. Among the clicked results, last Clicked result in the list appears more memorable than previous result. How user memorized rank of search results was studied to understand how user re-find the same result. The recalled rank differed from actual rank 33 % of time. Users correctly identified initial results 90 % of time and accuracy dropped as rank increased.

4.2 Recognizing Change in Result List

Most of the time user recognizes results that are different from initial result, but study showed that very different list can be recognized as the same if they maintain consistency in recalled aspect. To study how user recognizes as same or different when list is different form initial list it is constructed in following ways: (1) *Random Merge*: Four results viewed previously were randomly with top six results of new list. (2) *Clicked Merge:* Results clicked during session 1 were ranked first, followed by new results. The exact no of results preserved varied as a function of how many results were clicked. (3) *Intelligent Merge:* Old and new results were merged with an attempt to preserve the memorable aspects of the list during previous session. (4) *Original Merge:* The result list was exactly same as the originally viewed list. (5) *New:* The list was comprised of entirely new results. For intelligent merge user voted highest 81 % of time results returned by search engine were same.

4.3 Reusing the Search Results

Reusing the search result focuses on finding same result again which were found during previous session. To observe this user's history of clicking the result were captured by proxy in session 1 and during session 2 user had to re-finding and New-finding the results based on results visited during first session. For new finding tasks intelligent merge gains lowest mean and median task time, and for re-finding task original merge gains lowest mean and median task time. For re-finding task intelligent merge performs closest to original merge compared with other methods.

5 Click Through Bipartite Graph: [7]

Web search engine does not only retrieves the document relevant to query but it also ranks the documents such that most relevant document appears at the higher position of search result. When user is unable to formulate query to satisfy required information goal search engine provides query suggestion based on user's current query. With the help of click through bipartite graph we can efficiently measure query-document, document-document and query-query similarity. Document ranking can be assigned using click through bipartite graph using no of clicks on query URL pair. Highly clicked URL will be positioned on top of the SERP. Query suggestion is given based on overleaping URLs between queries in click through bipartite graph. Challenge with click through bipartite graph is that query-document relevancy is not calculated based on only no of click to the URL because we saw in previous section that due to positional bias higher ranked URL may get more click even if both URL contains same similarity to query. In order to avoid that multiple feature of web search log is taken into consideration (Fig. 1).

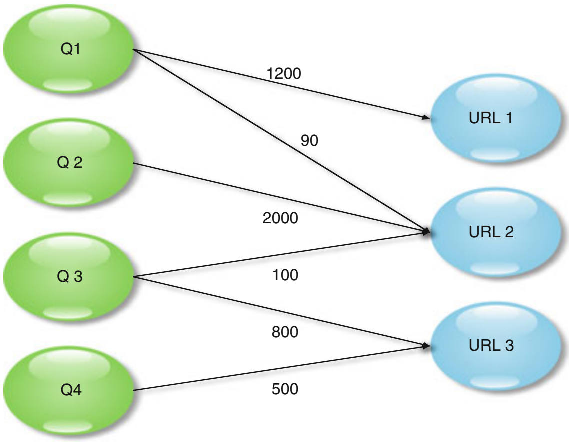

Fig. 1 Click-through bipartite shows query and url as nodes and link shows click rate [7]

Query document bipartite graph consist of triplet (q, d, t) where q denotes query d denotes document and t denotes no of times clicked. For considering multiple feature query and document are represented as vectors in query space Q_i and document space D_i where Q_i and D_i are sub spaces of Euclidian distance. Q_i and D_i may or may not be the same space.

Click through graph consists of query, URL and number of times URL clicked for a query. Relationship between query and URL can be identified by: (1) Euclidean distance [7]: it maps Query q and its feature in query space Q_i, and URL u and its features in URL space U_i. Based on Query q in Q_i and URL u in U_i Euclidean distance Di is calculated. (2) Co-visited method [8]: If two URLs are visited by same query it is possible that both URL presents similar information are called as co-visited. Number of times URL clicked for same query is used for calculating similarity between URLs. (3) Iterative Algorithm [8]: Iterative algorithm compared with co-visited method also checks for query similarity based on URLs visited by two or more queries. Based on derived relation from query and URL similarity, Iterative algorithm iteratively derives new relation which were not discovered in previous iterations. (4) Learning the query intent [9]: It is a semi-supervised method to learn query intent based on small amount of manually labeled queries and related URLs. When new query arrives which is not present in labeled queries it classifies URL to +/− based on labeled data.

6 Click Modeling [10]

User clicks are known to be biased based on position presented on result page. Main challenge is to model user click unbiasedly. Click model provides user search behavior which can be compared with judgements of web document and can be used in following ways: (1) *Automated ranking alteration:* Highly ranked results are altered with user preference to achieve user satisfaction. (2) *Search quality metrics:* Measuring user behavior and satisfaction using query-reformulation rate, abandon rate etc. (3) *Validation of judgement:* User feedback as explicit judgement is compared with implicit judgement provided by domain expert. (4) *Online advertisement:* Based on user's search history feature clicks are predicted and mapped as advertisement to increase revenue.

Click modeling can be done using: (1) Positional model [10]: it calculates probability of URL being clicked using relevance of query and URL and position at which URL is presented in result page. (2) Cascade Model [10]: It assumes that user examines the URL sequentially from top to bottom and stops as soon as relevant information is found. Click on ith document means: (1) URLs above ith position are skipped by user and are not relevant to query. (2) Ith URL is relevant to query. (3) Dynamic Bayesian Network [10]: It represents user click into three variables E_i: user examined the URL, A_i: user attracted by the URL, S_i: user satisfied by the result? If answer for variable is yes then it takes value as 1 otherwise 0. It says: (1) if user is attracted and examines the URL then Click of URL occurs. (2) Attraction dependent on URL's relevancy to query. (3) User examines abstracts of results from top to bottom and clicks on URL with certain probability of being satisfied. (4) As soon as user is satisfied by ith URL, probability of examining URL ith below is 0. (4) Click Chain Modeling (CCM) [11]: in CCM each URL on result page contains its own probability of examination, probability of click and relevance to query. Main difference here is probability of current URL is connected with all previous URLs and their probability.

6.1 Context Aware Ranking in Web Search [12]

Context of a search query gives meaningful information about search intent. For example user raises query "apple" after searching "Samsung mobile phone", it is very likely that user is searching for apple mobile phone rather than apple fruit. There are two main challenges in context aware ranking: (1) Using context to rank result. (2) Using different types of contexts such as user query, URL clicked. Context aware ranking principles: (1) Reformulation: user reformulates the query to be more specific about information needs. (2) Specialization: User issues specialized query to see results that are more specific about user's intent. (3) Generalization: User may ask a query more general than previous one to gain more general knowledge. (4) General association: When query is generally associated with its context, context may help to narrow down the user's search intent.

7 Short Term and Long Term User Interests

This section contains: (1) Predicting short term user interests. (2) Using Long term search history to improve search accuracy.

7.1 Predicting Short Term User Interests: [13]

Short term search is limited to only single search session containing consecutive queries. Query context is pre-query activities that includes previous query and page visited last in past. Developing and evaluating user interest model for current query, its context and their combination is called as intent. Based on past query and URL visited within session context is created. When user issues new query Q3, intent is constructed from context for current query Q3 and based on that intent optimal result is returned to user. Short term user interests can be modeled in: (1) Query model: Open Directory Project (ODP, dmoz.org) provides human-edited directory of web. It contains categorized list of URLs. In this model categories for top 10 result are retrieved from dmoz.org. Based on URL clicks of user their categories are mapped and user preference for particular category is saved. When same query is again raised by user search engine assigns higher ranks to URLs that belongs to category which user previously visited. (2) Context model: Context model also categorizes URLs visited by user in categories provided by ODP. Weight to particular category is assigned based on dwell time on the page. If user visits URL for more than 30 s than that URL contains interesting contents. (3) Intent model: Intent model is combination of query model and context model. Since query model includes information from current query and context model includes information about user's activity in current session, combination of both information gives more accurate results. (4) Relevance model or Ground truth: Relevance model predicts future actions. It assigns higher weights for most recent actions of user. This model captures most recent user action as more valuable for constructing context. This model generates best result based on observation that each user action leads closer to information goals.

7.2 Long Term Search History to Improve Search Accuracy: [14]

Most existing retrieval system including search engine offers generalized web search interface which is optimal for all web users. Retrieval of document is made based on only the query submitted by user and ignoring user's preferences or the search context. When user submits query "python" is ambiguous and search result may contain mixed content which is non-optimal for the user. Instead of using query only as retrieval option, user search context can be used to match with user's intended information needs. There is wide variety of search contexts like bookmarks, user interests in particular categories, user's long term search history etc. In long term search history logs of user's search history is maintained based on URL clicked by user. For example user has searched for "debugging" and "Java code" and currently searching for "python" suggests that user is searching for python related to programming context. Second optimization can be done based on user's past searches for example if user searched for "Perl programming" in past and visited some web pages and if same search is repeated then based on user's past visited URL current SERP can be re ranked based on user preference.

8 Search Trail and Popular URLs

This section contains: (1) Search trails. (2) Evaluating effectiveness of search trails. (3) Using Popular URLs to enhance web search.

Table 1 Shows session example

Time	Action	Value
t0	Query	Buy Samsung mobile online
t1	SERP click	http://www.samsung.com
t2	URL click	http://www.samsung.com
t3	Query	Samsung galaxy s6 edge review
t4	SERP click	http://www.in.techradar.com
t5	Query	Buy Samsung galaxy s6 edge
t6	Re-query	Buy Samsung galaxy s6 edge flipkart
t7	SERP click	http://www.flipkart.com/
t8	URL click	https://www.flipkart.com/checkout/

8.1 *Search Trail [15]*

User with certain information goal submits query to search engine and visits URL presented on result page. User also visits URLs that are presented on web page whose URL address is returned by search engine. Search trail consists of both URL presented on result page and URL that were not presented on result page. As shown in session example in Table 1 search trail for "buying a Samsung mobile phone" at time t1 clicks on URL presented on result page and at time t2 user clicks on link from web page visited at time t1 which was not presented on result page. At time t3 based on information gained from URL visited at time t1 and t2 user formulates new query at time t3 and finally search trail ends when user performs payment of buying a mobile at time t8. Search trail may span to multiple session.

8.2 *Evaluating Effectiveness of Search Task Trails [16]*

Experiment conducted to measure effectiveness of search task trails on large scale dataset from commercial engine shows results that (1) User tasks trails are more accurate as compared to session and query modeling. (2) Task trails provides unambiguous user information needs. (3) Task trail based query suggestion performs well as compared with other models. Query task clustering approach: Queries which belong to same task can be combined into a single cluster. Based on observation consecutive query pairs are more likely belong to same task rather than non-consecutive ones. User search interests are measured by using mapping URL belongs to same task into categories provided by ODP on dmoz.org. Dwell time, Hidden Markov Model to measuring success rate of search and number of clicks on URL are taken as user implicit feedback. Query suggestion models: (1) Random walk, (2) Log likelihood, (3) co-occurrences is used to measure the performance of task trails modeling.

8.3 Using Popular URLs to Enhance Web Search: [6]

Query suggestion offers similar query to current query of user. Query suggestion allows user to express query more specifically leading to improved retrieval performance. Search engines gives query suggestion based on query reformulation of users. Examining the most common URLs visited by majority of user for a given query is referred as popular URLs. Popular URLs may not be ranked in result list of query or may not contains words similar to user query. This approach gives user a shortcut to reach information goal. To examine the usefulness of destinations four system were used in study: (1) baseline web search system with no explicit support for query recommendation, (2) A search system with a query suggestion method that recommends additional query (3) query destination which suggest popular URL destination for given query. (4) Session destination which suggests endpoint of session trails. Among four system query destination achieves highest positive user feedback and mean average time to complete task was minimum.

9 Summary

This paper provides survey on search log mining for web search, with focus on accurately interpreting user feedback and various methods to model user's implicit feedback. By modeling user's implicit feedback search engine results can be re-ranked to improve retrieval quality of search engine, query suggestion and URL recommendation.

References

1. Broder, A.: A taxonomy of web search. In: ACM Sigir forum, vol. 36, no. 2, pp. 3–10. ACM (2002)
2. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 154–161. ACM (2005)
3. Radlinski, F., Kurup, M., Joachims, T.: How does clickthrough data reflect retrieval quality? In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 43–52. ACM (2008)
4. Downey, D., Dumais, S., Liebling, D., Horvitz, E.: Understanding the relationship between searchers' queries and information goals. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 449–458. ACM (2008)
5. Agichtein, E., Brill, E., Dumais, S., Ragno, R.: Learning user interaction models for predicting web search result preferences. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3–10. ACM (2006)
6. Teevan, J.: How people recall, recognize, and reuse search results. ACM Trans. Inf. Syst. (TOIS) **26**(4), 19 (2008)

7. Wu, W., Li, H., Xu, J.: Learning query and document similarities from click-through bipartite graph with metadata. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 687–696. ACM (2013)
8. Xue, G.-R., Zeng, H.-J., Chen, Z., Yu, Y., Ma, W.-Y., Xi, W.S., Fan, W.G.: Optimizing web search using web click-through data. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, pp. 118–126. ACM (2004)
9. Li, X., Wang, Y.-Y., Acero, A.: Learning query intent from regularized click graphs. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 339–346. ACM (2008)
10. Chapelle, O., Zhang, Y.: A dynamic bayesian network click model for web search ranking. In: Proceedings of the 18th International Conference on World Wide Web, pp. 1–10. ACM (2009)
11. Guo, F., Liu, C., Kannan, A., Minka, T., Taylor, M., Wang, Y.-M., Faloutsos, C.: Click chain model in web search. In: Proceedings of the 18th International Conference on World Wide Web, pp. 11–20. ACM (2009)
12. Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E., Li, H.: Context-aware ranking in web search. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 451–458. ACM (2010)
13. White, R.W., Bennett, P.N., Dumais, S.T.: Predicting short-term interests using activity-based search context. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1009–1018. ACM (2010)
14. Tan, B., Shen, X., Zhai, C.X.: Mining long-term search history to improve search accuracy. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 718–723. ACM (2006)
15. Bilenko, M., White, R.W.: Mining the search trails of surfing crowds: identifying relevant websites from user activity. In: Proceedings of the 17th International Conference on World Wide Web, pp. 51–60. ACM (2008)
16. Liao, Z., Song, Y., He, L.-W., Huang, Y.: Evaluating the effectiveness of search task trails. In: Proceedings of the 21st International Conference on World Wide Web, pp. 489–498. ACM (2012)