Are Search Engine Users Equally Reliable?[1]

Qianli Xing, Yiqun Liu, Rongwei Cen, Min Zhang, Shaoping Ma, Liyun Ru

State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University, Beijing, China, 100084

xingqianli@gmail.com

ABSTRACT

In this paper, we study on the reliability of search engine users using click-through data. We proposed a graph-based approach to evaluate user reliability according to how users click on search result lists. We tried to incorporate this measure of reliability into relevance feedback for improving ranking performances. Experimental results indicate that the proposed approach is both effective and applicable.

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

General Terms: Algorithms, Measurement, Experimentation

Keywords: User reliability, Web search, relevance feedback

1. INTRODUCTION

Relevance feedback techniques are now widely used in search engines to improve both retrieval quality and user experience. But as far as we know, most feedback models based on click-through information don't consider much about the quality of feedback sources, saying the user reliability in case of Web search. Sadagopan et al. [1] identified typical and atypical user sessions in click streams, but it's on session level and didn't distinguish different clicks. We can imagine that different users are of different level of expertise due to various reasons such as profession and education background. For example, we believe that expert users click better results than beginners in most cases on search engine result page, which means that feedback data from different users are of different significance. So relevance feedback may be not accurate enough if we don't separate different users in feedback models. Therefore, reliability of users should be evaluated when we collect feedback information.

In this paper, we propose a graph-based approach to measure user reliability. First we create a graph using click-through data, describing the relation among users by clicks. Then we use an iterative algorithm to compute reliability scores on the graph. And in latter experiment we observe the influence of the reliability score to performance of relevance feedback. Results show that feedback performance indeed varies from user to user according to the reliability score and a tendency can be figured out. It convinces our idea that user reliability should be considered if we want to get better results in relevance feedback.

2. EVALUATING USER RELIABILITY

2.1 Definition

In order to evaluate reliability, we should first give definition of user reliability. In Web search, we regard the reliability of a user to be the ability that he or she finds answers for a query. Generally in practice, we regard a search engine user to be reliable if he/she is able to click relevant results with a high probability under different queries. It looks like that we can identify reliable users by checking relevance of clicked results. But unfortunately, it's almost impossible to know which results are relevant for most queries without manual efforts. There are too many unpredictable queries and we can't perform manual relevance judgments for all of them. In this case, an intuitive assumption is that results clicked by most users tend to be relevant. But this assumption is weak [2] so in this paper we try to indentify reliable users using other information rather than the known relevant results. Here gives our definition of search user reliability.

DEFINITION: A search engine user is reliable if most results he /she clicked are agreed by a sufficient number of other users.

It means the reliability could transfer between two search users if they agree on some result selections. So we try to identify user reliability according to user relationships.

2.2 Graph Model

We now propose a graph model to measure the reliability defined above. Firstly we create a graph called user-click graph to reflect relationships among users by their clicks. In the graph, each node represents a unique search engine user. Each edge between two nodes represents that the corresponding two users both click on a same result while questioning a same query. All edges in graph are bidirectional and there could be multiple edges between two nodes. Besides, each node has a property which records the submitted queries and the number of clicks under each query of the corresponding user.

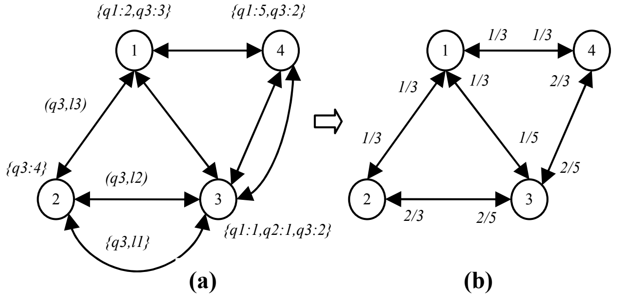

Figure 1. A sample of user-click graph

Figure 1(a) is an example showing what this graph looks like. User 2 in the graph, for example, clicked link *l1,l2,l3,l4* under query *q3*. User 1 also clicked *l3* under *q3* so they are connected by

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

[1] Supported by Natural Science Foundation (60736044, 60903107) and Research Fund for the Doctoral Program of Higher Education of China (20090002120005).

an edge. Similarly, user 3 had two common clicks with user 2 so there are 2 edges between them. Figure 1(b) is another form of the graph which merges multiple edges between two nodes into one and assigns two weights, for two directions respectively, to the edge. So we design an iterative algorithm to compute reliability score for each user. The reliability score r_i of user *i* is defined as:

$$r_i = \sum_{(j,i)\in E} P_{ij} W_{ij} r_j$$

Where *E* is the set of edges, $P_{i,j}$ is the penalty factor for user *i* with user *j* and $W_{i,j}$ is the weight of the edge from user *j* to user *i*.

$$P_{ij} = \underset{q\in Q_{ij}}{Avg} \frac{|D_i(q)|}{|C_i(q)|} \qquad W_{ij} = \frac{|E_{ji}|}{indegree(i)}$$

Q_{ij} is the set of common queries for user *i* and user *j*. $D_i(q)$ is the set of results only clicked by user *i* under query *q* and $C_i(q)$ is the set of results clicked by user *i* under query *q*. E_{ji} is the set of edges between user *i and j*. Note that P_{ij} is different from P_{ji}, so is W_{ij} and W_{ji} .

Taking the graph in Figure 1 for example, for user 2, we have r_2=(3/4)•(r_1/3+2r_3/3). Link *l4* is not clicked by any other user besides user 2 so there is a penalty factor of 3/4 for user 2.

This iterative algorithm can be proven to converge. We initialize r_i with value 1/*N* and stop after 20 iterations.

3. EXPERIMENT AND EVALUATION

3.1 Dataset

We perform our approach on practical search log of one day (1st Oct 2009) from a major Chinese search engine called *Sogou.*

When generating the graph, we eliminate users who perform no more than 3 clicks because these users rarely use the search engine and it's not so helpful to evaluate their reliability. About 80% users are eliminated after this process.

Click distribution (CD)[3] describes the density of clicks for a given query. For queries with CD value above 0.5, which have converged clicks on some result respectively, we remove them from the dataset. The reason for this is that most of these queries (hot navigational queries for example) have obvious answers so they are not discriminative to tell which user is reliable. Involving these queries will cost extra running time but in fact search engine has performed quite well on these queries so we don't concern them as much as other queries in this feedback problem.

After these data preprocesses, there are finally 691,290 users and 3,382,001 unique clicks in the graph.

3.2 Effectiveness of Reliability

We apply the proposed iterative algorithm on the generated graph to compute reliability score for each user.

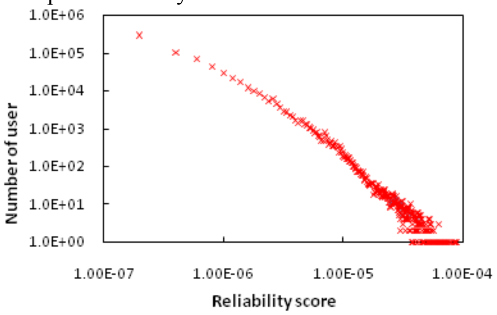

Figure 2. Distribution of reliability score

After taking log of both scales, the distribution of user number on reliability score basically fits the power-law which is shown in Figure 2. The distribution accords with our expectation that the number of user decreases rapidly as reliability increases.

In order to evaluate the effectiveness of our approach, we generated the list of users in descending order of their reliability scores, and we segmented it into 10 buckets. Each bucket has same number of users. So the first bucket contains 10% users with the highest reliability scores and the 10th bucket contains 10% users with the lowest reliability scores.

We let users in each bucket vote for the result relevance. One click stands for one vote, and the voting result represents the relative relevance judgment by the voted users. We evaluate the performance of feedback for each bucket of users on our test dataset, in which the relevance of each top 10 results for 600 randomly selected queries are manually labeled with one of the four grades(not relevant, little relevant, relevant, strong relevant).

We measure the feedback accuracy by Kendall's tau, which describe the consistency of pairwise relevance judgment between the labeled result and the feedback result. Figure 3(a) shows the performance of each bucket of users.

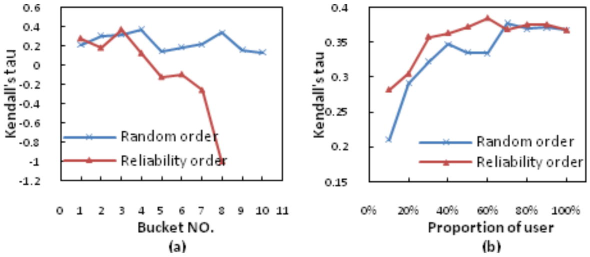

Figure 3. Performance of relevance feedback

We can see the tendency that users with higher reliability scores obtain better relevance feedback result than those with lower scores. It shows that the reliability score obtained by the proposed algorithm represents the reliability of search users.

With the tendency, we are interested in knowing how many users provide best feedback accuracy. Figure 3(b) shows the feedback performance when we adopt different proportion of users from top reliability. It indicates that feedback performs better by involving more users at the beginning and performs best with top 60% users. Users with high reliability scores also perform better than those randomly chosen ones.

4. CONCLUSION AND FUTURE WORK

The main contribution of our work is that we proposed an approach to identify reliability of search users, which can be further adopted in various feedback models and personalized search services. Experimental results show that relevance feedback from reliable users is more accurate than that from relatively unreliable users. In the future work, it is interesting to try incorporating the reliability score into some feedback model to achieve higher accuracy for relevance feedback.

5. REFERENCES

[1] N. Sadagopan and J. Li. Characterizing typical and atypical user sessions in clickstreams. In *WWW 08 :Proceedings of the 17th international conference on World Wide Web*

[2] R. Cen, Y. Liu, M. Zhang, L. Ru and S. Ma. Automatic Search Engine Performance Evaluation with the Wisdom of Crowds. In *AIRS 2009, Japan*

[3] U. Lee , Z. Liu , J. Cho, Automatic identification of user goals in Web search, In *WWW 05: Proceedings of the 14th international conference on World Wide Web*