Automatic Search Engine Performance Evaluation with Click-through Data Analysis*

Yiqun Liu, Yupeng Fu
State Key Lab of Intelligent
technology & systems
Tsinghua University
Beijing, China P.R.

liuyiqun03@gmail.com

Min Zhang, Shaoping Ma
State Key Lab of Intelligent
technology & systems
Tsinghua University
Beijing, China P.R.

{z-m, msp}@tsinghua.edu.cn

Liyun Ru
Technology Research &
Development center
Sohu Incorporation
Beijing, China P.R.

ruliyun@sohu-rd.com

ABSTRACT
Performance evaluation is an important issue in Web search engine researches. Traditional evaluation methods rely on much human efforts and are therefore quite time-consuming. With click-through data analysis, we proposed an automatic search engine performance evaluation method. This method generates navigational type query topics and answers automatically based on search users' querying and clicking behavior. Experimental results based on a commercial Chinese search engine's user logs show that the automatically method gets a similar evaluation result with traditional assessor-based ones.

Categories and Subject Descriptors
H.3.4 [**Systems and Software**]: Performance evaluation

General Terms
Experimentation, Measurement

Keywords
Performance evaluation, Click-through data analysis

1. INTRODUCTION
Evaluation is one of the key questions in information retrieval (IR) research. Currently most IR evaluation researches are based on Cranfield's studies in [2]. A Cranfield-like approach is based on a set of query topics, their corresponding answers (usually called qrels) and evaluation metrics. Queries are processed by an IR system and results are compared with qrels using evaluation metrics. The annotation of correct answers is usually the most difficult part in this kind of evaluations. The time cost of commercial search engine evaluation would be unacceptable using manual annotation. Therefore an automatically evaluation process which is objective, reliable and up-to-date will be useful for search engines to monitor and improve search performances.

Several recent attempts have been made towards automatically evaluation in order to tackle the difficulties related to manual assessment. However, researchers' efforts haven't involved the use of user behavior information which is valuable for IR evaluations until Joachims [3] designed an unbiased automatic assessment using the click-through data. But Joachims' work was based on small scale data collection (only several hundred user clicks are collected) so the reliability and effectiveness of using click-through data in IR evaluation remains to be studied into.

The main contributions of our work are that we propose a fully

automatic approach that measures a search engine's performance based on click-through data. Instead of building a query set and annotating relevant documents manually, we select topics and annotate answers automatically by analyzing users' query log and click-through data. The correctness of both the automatically annotated answers and the evaluation results are compared with manual ones to verify the reliability of this approach.

2. EVALUATION BASED ON CLICK-THROUGH DATA ANALYSIS
Search engine performance should be evaluated by its effectiveness in meeting different kinds of information needs. Based on the query log analysis of Alta Vista, Broder [1] groups search requests into 3 categories: navigational, informational and transactional. The major difference between navigational type queries and the other two types of queries is whether the user has a fixed search target page or not. We focus on navigational type queries in our evaluation approach because: Firstly, users always need to find the URL of a particular Web site or a certain Web page. Navigational queries are so frequently proposed that many commercial search engines design specific gateways for them such as Google's "I'm feeling lucky" function. Secondly, there is usually only one correct answer for each navigational query so it avoids the problem of missing correct answers. This problem may lead to failure in evaluation because it causes lower estimation of the search engines which provide answers that are not annotated.

2.1 Feature Extraction and Navigational Type Query Selection
Because only navigational queries are adopted in our approach, we need to pick up such kind of queries using click-through data analysis. In our previous work [5], we classify user queries according to their click-through information. With a learning-based algorithm, over 80% queries were classified correctly. This classification result was obtained using a training set of 198 queries and a separate test set of 233 human annotated queries. The algorithm was used in our evaluation approach to separate navigational type queries from the other ones.

2.2 Automatically Answer Annotation
Click distribution is a feature proposed by Lee et al in [4]. It was used by Lee in query type identification but we found that its variation can also be used in the automatically annotation process. Click distribution (CD) of a query q is defined as:

$$CD(Query\ q) = \frac{\#(Session\ of\ q\ that\ involves\ clicks\ on\ R_{most})}{\#(Session\ of\ q)} \quad (1)$$

For a navigational type query q, R_{most} is defined as the URL which is clicked by the most Web search users who are querying q. Users who propose a navigational type query will click a certain result because they consider this result as their search target. Hence R_{most} is likely to be the correct answer for q as long as search engines can return the answer at a relative front position so that the users can find and click it. This gives the possibility of annotating search target pages for navigational queries using the click-through information of different users. Given the definition:

$$ClickFocus(Query\ q, Result\ r) = \frac{\#(Session\ of\ q\ that\ clicks\ r)}{\#(Session\ of\ q)} \quad (2)$$

The following equation holds according to (1) and (2):

$$ClickFocus\ (Query\ q, R_{most}) = CD(Query\ q) \quad (3)$$

Then the annotation process can be described as a process to locate R_{most} for each Query q. The follow-up steps of the approach including search engine result crawling and evaluation metrics calculation are similar with traditional Cranfield-like approaches.

3. EXPERIMENTS AND DISCUSSIONS

All the experiments are based on the click-through data collected by Sogou.com from June, 2006 to January, 2007. Sogou.com is one of the largest commercial search engines in Chinese Web environment. Part of the click-through data is online for free download at http://www.sogou.com/labs/dl/q.html. The log records about 1.5 million querying or clicking events per day.

In contrast to the time-consuming and labor intensive manually assessment, our approach is efficient and it can process about 400 topics automatically in merely one hour. This efficiency result was obtained using one Pentium 4 PC (costs about 800 US dollars) in 100M LAN network environment.

3.1 Answer Annotation Experiment

We annotated three groups of queries using click-through logs during different time periods. About 5% of the annotated queries are picked up randomly and manually checked for correctness.

**Table 1 Size of the annotated query set and accuracy of the annotated answers**

	#(Annotated queries)	#(Checked sample set)	Accuracy
Jun. 06 - Aug. 06	13,902	695	98.13%
Sept.06 - Nov. 06	13,884	694	97.41%
Dec. 06 - Jan. 07	11,296	565	96.64%

According to the results shown in Table 1, we can see that during each time period, over ten thousand queries are successfully annotated and over 95% of the sampled annotated answers are correct. The size of the 3rd annotated query set is a bit smaller because the time period is shorter than the other two periods.

We checked the wrongly-annotated answers in the sample set and found that these answers are usually sub-sites of the correct answers instead of the correct answers themselves. This is caused by the fact that the correct answers are always queried because of their sub-sites. For example, http://mail.163.com/ is more welcomed by users than its homepage because it is one of the most popular E-mail service providers in China. Therefore, when users propose the query of "163" they usually click the email service portal and the automatically approach annotate it as the answer instead of the homepage (http://www.163.com/).

3.2 Performance Evaluation Experiment

With the query set and the answer set annotated in Section 2, search engine performance can be evaluated using traditional metrics. MRR [1] are used in our experiments for evaluating navigational type queries and results are shown in Figure 1.
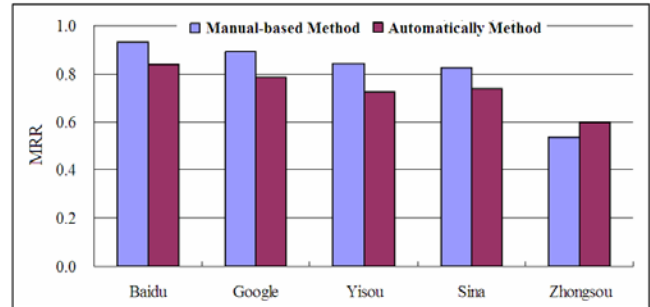
Figure 1 Comparison in five search engines' evaluation results between manual-based method and automatically method

In manual-based methods, a set of 320 navigational queries are selected randomly from the query logs. The answers are annotated by assessors using pooling method and the result pool is built aggregating search results collected from the search engines to be evaluated. According to Figure 1, we found that the automatically evaluation result has the same performance ranking as the manual one. The correlation value between MRRs of the two methods is 0.965, which indicates the two evaluation results are quite similar.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed an automatically performance evaluation approach for Web search engines. This approach is based on users' click-through behavior which is recorded in search engines' user logs. The method is a Cranfield-like one and it constructs navigational type query set and annotates answers automatically. According to experimental results, most of the answers are annotated correctly. Besides, automatically evaluation results are highly correlated with manual-based ones.

Future study will focus on the following aspects: How much click-through data is needed for an evaluation which is both reliable and efficiency? Whether the evaluation results will be more reliable with the help of click-through data collected from two or more search engine and how to combine the information? Besides those questions, we also want to find out whether the automatically method can be extended to the evaluation of other types of queries besides navigational type ones.

5. REFERENCES

[1] A. Broder, A taxonomy of Web search. SIGIR Forum Volume 36 Number 2, 2002.

[2] T. Saracevic, Evaluation of evaluation in information retrieval, Proceedings of the 18th ACM SIGIR, 1995.

[3] T. Joachims, Evaluating Retrieval Performance Using Click-through Data. In SIGIR Workshop, 2002.

[4] U. Lee, Z. Liu and J. Cho, Automatic Identification of User Goals in Web Search, in the 14th WWW Conference, 2005.

[5] Y. Liu, M. Zhang, L. Ru and S. Ma, Automatic Query Type Identification Based on Click-through Information, in LNCS 4182, pp. 593–600, 2006.

---

[1] Mean Reciprocal Rank (MRR) is a metric in navigational type evaluation. RR equals to the reciprocal of the correct answer's ranking in the result list and MRR is the mean of the topics' RRs.