Weighted Information Gain and User Clicks on Web Search Results

Yun Zhou , W.Bruce Croft,

Department of Computer Science, University of Massachusetts Amherst,
140 Governor Dr. , Amherst.MA,01002, USA
{yzhou, croft}@cs.umass.edu

Abstract. In this poster, we demonstrate that the WIG (Weighted Information Gain) technique, originally proposed for retrieval performance prediction and shown to be effective particularly in Web search environments, has an interesting connection with user clicks on Web search results. Specifically, we observe that high WIG scores generally suggest more clicks. This makes WIG a useful feature for predicting user's preference for search results, which has potential applications in many important areas such as the automatic tuning of search engine parameters, personalization, sponsored search and others.

Keywords: WIG, click, prediction, user preference

1 Introduction

WIG (Weighted Information Gain) was demonstrated as an effective technique for retrieval performance prediction in Web search environments [1]. A significant correlation between the WIG score and retrieval performance was observed in various Web search scenarios. However, the evaluation of WIG in [1] was performed under laboratory settings which consist of carefully-selected topic sets with relevance judgments made by human assessors and pre-defined retrieval tasks. The high cost of producing relevance judgments makes it difficult to test the performance of WIG in real-world Web search. Instead of relying on human relevance judgments, we make use of user clicks on search results to approximate relevance judgments. In fact, the click of a document can be viewed as an implicit and rough relevance feedback made by the user. Although it can be noisy, click information, when used in aggregation, is a good indicator of relevance [2]. In this poster, we experiment on realistic Web data with click information gathered from a commercial search engine. Consistent with the finding in [1] that high WIG scores generally correspond to high retrieval performance, we observe a tendency that high WIG scores predict more clicks on search results. This confirms earlier experiments showing that WIG is a useful feature for predicting the outcome of a query, which has potential applications in many important areas such as the automatic tuning of search engine parameters, personalization, sponsored search and others.

2 WIG (Weighted Information Gain) Score

The WIG technique was first introduced in [1] for predicting query performance. We briefly revisit how the WIG is calculated for a given query.

Specifically, given query Q, a ranked list L of documents in response to Q, and a collection C , the WIG score is computed as follows:

$$WIG(Q,C,L) = \frac{1}{K} \sum_{D_t \in T_K(L)} \sum_{\xi \in F(Q)} \lambda_\xi \log \frac{P(\xi \mid D_t)}{P(\xi \mid C)} \qquad (1)$$

where $P(\xi|D_t)$ denotes the probability that feature ξ will occur in Dt, $P(\xi|C)$ denotes the probability that feature ξ will occur in collection C, F(Q) consists of a set of features expanded from the original query Q, K is a cutoff rank and TK(L) contains the top K documents in L, λ_ξ is a normalization parameter and is set as follows:

$$\lambda_\xi = \begin{cases} \frac{\lambda_T}{\sqrt{|T(Q_i)|}}, \xi \in T(Q_i) \\ \frac{1-\lambda_T}{\sqrt{|P(Q_i)|}}, \xi \in P(Q_i) \end{cases} \qquad (2)$$

where $|T(Q_i)|$ and $|P(Q_i)|$ denote the number of single term and proximity features in $F(Q_i)$ respectively.

The WIG score can be viewed as the difference of the average language-model score over the top ranked documents between the actual ranked list and a random ranked list. A strong positive correlation between WIG scores and retrieval performance was observed in [1].

3 EXPERIMENTAL DESIGN AND RESULTS

We first introduce the dataset used in this poster. It is a query log file that contains about 149 million queries collected by a Web search company during of one month period (from May 1 to May 31 of 2006). For each query, the following information is associated: (1) query ID, (2) the time the query is submitted, (3) the content of the query, (4) the result(s) clicked by the user who submits the query, (5) the number of returned results. Notice that (4) and (5) are not available if no results are clicked for the query. No relevance information on queries is available. This dataset represents the kind of information that a typical web search engine can readily obtain from user interaction.

Our hypothesis is that the higher the WIG score is, the more search results the user will click on. That is, the WIG score is positively correlated to user's preference for search results. We next describe our experimental design to test the above hypothesis. We randomly sample 2000 queries from the query log file. We assume that each query is issued by a unique user (that is, one-to-one correspondence between a query and a user) .We divide these queries into three groups according to the number of

results the user clicked when seeing the ranked list of results in response to her query. Table 3.1 gives the details. For example, Group A represents those users who do not click any of the returned results. In fact, group A, B and C represent three levels of user interest in search results in ascending order. The percentage of each group in our sampled query set is also provided in Table 3.1. We can see that the majority of users only click one of the retrieved results.

Table 3.1. Division of test queries into three groups based on the number of clicked results

Group	A	B	C
# of clicked results	0	1	>=2
Percentage	34.8%	50.1%	14.1%

Table 3.2. Distributions of WIG scores for Group A, B and C

Group	A	B	C
Sample mean of WIG score	5.340	6.040	6.648
Sample Variance of WIG score	11.645	8.120	8.877
Size	695	1002	280

For each query in the sampled query set, we compute the WIG score. For WIG calculation, in addition to the query itself we need the ranked list in response to the query and a collection. We use the provided search engine API to download the top ranked documents for the query. The GOV2 collection is used to approximate the Web collection statistics required in WIG calculation. The parameter settings of WIG are the same as used in [1]. The distributions of WIG scores for the three groups are presented in Table 3.2. We adopt two statistics to represent the distribution of WIG scores for each group: sample mean and sample variance. The size of each group (the number of queries in the group) is also provided.

Let WIG(A),WIG(B) and WIG(C) represent the mean of WIG scores in group A, B and C respectively, that is,

$$WIG(A) = \frac{1}{|A|}\sum_{q \in A} WIG(q) \qquad (3)$$

Where |A| is the size of group A and WIG(q) is the WIG score for query q. WIG(B) and WIG(C) have similar definitions. From Table 3.2 we observe that WIG(C) > WIG(B) and WIG(B) > WIG(A). Further investigation shows that both of the two differences (WIG(C) - WIG(B) and WIG(B) - WIG(A)) are statistically significant at the 95% confidence level according to the t test [3]. This shows that high WIG scores suggest more clicks, which is consistent with the previous finding that high WIG scores generally correspond to high retrieval performance.

We want to point out the correlation between WIG and clicks is not strong enough to make a conclusion that WIG alone can accurately predict clicks, considering the large variance of WIG scores in each group as shown in Table 3.2. This is due to the fact that user preferences for search results depend on many factors other than retrieval quality. For example, educational background may have a large impact on user preferences. In fact, we observe in the dataset that, for the same query, some users do not click on any of the returned results while others do click. Since WIG is an effective feature only for predicting relevance and clicks are only related to relevance to some degree, we do not expect that clicks can be accurately predicted by WIG alone. More experiments are needed to study how WIG can be combined with click-based relevance models.

Acknowledgments. This work was supported in part by Microsoft and Google.

References

[1] Y.Zhou, W.B.Croft. Query performance prediction in Web search environments . In *the Proceedings of the 30th Annual International ACM SIGIR Conference (SIGIR 2007),* 543-550, (2007)

[2] T.Joachims, L. A. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *the Proceedings of the 28th Annual International ACM SIGIR Conference (SIGIR 2005),* 154–161, (2005).

[3] George Casella and Roger L. Berger, *Statistical Inference*, Duxbury, (2002).