Determining User's Interest in Real Time

Sanasam Ranbir Singh, Hema A. Murthy, and Timothy A. Gonsalves
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai-600 036, India
{ranbir,hema,tag}@lantana.tenet.res.in

ABSTRACT

Most of the search engine optimization techniques attempt to predict users interest by learning from the past information collected from different sources. But, a user's current interest often depends on many factors which are not captured in the past information. In this paper, we attempt to identify user's current interest in real time from the information provided by the user in the current query session. By identifying user's interest in real time, the engine could adapt differently to different users in real time. Experimental verification indicates that our approach is encouraging for short queries.

Categories and Subject Descriptors

H.3.3 [**1**]: Information Search and Retrieval—*Search process*

General Terms

Measurement, Experimentation

Keywords

Users clicks, search engine, users' interest, real time

1. INTRODUCTION

Most of the currently adopted Search Engine's (SE) optimization techniques such as personalization [5], query expansion [2], user's intent [3, 4] try to predict user's interest from past information collected from different sources such as user's predefined interest, user's previously submitted queries and corresponding clicks etc. But, user's current interest for the same query may be different at different times, different places. Considering the famous disambiguation problem for the query `jaguar`, it is often difficult to predict whether user is looking for a `car` or a `cat`. An interesting question that arises is *'is it possible to learn user's current interest from the available information of the current query session, learn the user's interest and optimize the results instantly in real time?'*. In our study, a query session starts when user submits a new query and it terminates when user changes the query or terminates the search. If a user clicks on few links related to `car` (query being jaguar), it clearly indicates that he/she is currently looking for the `jaguar car`. If user further expands the search by extending to next page or reformulates the query, then the SE can instantly optimize the results by selecting only `car` related results or reformulating the query. Optimizing the results in real time can provide a unique experience to the users and make the search engine self adaptive. This is the main motivation of this paper.

The studies on search results clustering [6] have conceptual similarity in which similar results are placed in groups and a user can explore the groups based on his interest. But, learning user's current interest and user's current search behaviour in real time has a potential for the search engines to adapt differently to different users in real time. In this paper, we focus only on learning user's current interest from the clicked information submitted by the user at the time of search. The model can also merge with the information learnt from the past to enhance the performance, but such a study is beyond the scope of this paper.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. DETERMINING USER'S INTEREST

We define user's interest as a set of terms that can well represent the user's interest. A term could be an `unigram` or `bigram`. In the above `jaguar` example, the set of terms such as {`car, model`} could represent user's interest that user is looking for the jaguar car. We now formally define the problem as follows.

Problem Statement: Let $\mathcal{T}$ be the set of results that has been exposed to the user and each result t_i in $\mathcal{T}$ be a set of terms present in the snippet of the i^{th} result. We divide the set $\mathcal{T}$ into two disjoint subsets $\mathcal{T}_c$ and $\mathcal{T}_{\overline{c}} = \mathcal{T} - \mathcal{T}_c$ representing the set of clicked and not clicked results respectively. Let $\mathcal{W} = \cup_i^n t_i$ be the set of terms in $\mathcal{T}$, where $n = |\mathcal{T}|$. Now the problem is to determine a set of terms $\omega \subseteq \mathcal{W}$ which can well discriminate $\mathcal{T}_c$ and $\mathcal{T}_{\overline{c}}$ such that terms in ω frequently occur in $\mathcal{T}_c$ and less frequently occur in $\mathcal{T}_{\overline{c}}$. If $P_c(w)$ and $P_{\overline{c}}(w)$ are the probability distribution of w over $\mathcal{T}_c$ and $\mathcal{T}_{\overline{c}}$ respectively, we define a weight on each word w as

$$d(w) = \mid P_c(w) - P_{\overline{c}}(w) \mid . \log \frac{(2 - P_{\overline{c}}(w))}{(2 - P_c(w))} \qquad (1)$$

The values of $d(w)$ ranges from [-1,1]. Larger the value of $d(w)$, higher is the probability of occurring w in $\mathcal{T}_c$. Large value of $\mid P_c(w) - P_{\overline{c}}(w) \mid$ in Equation 1 means high frequency of occurrence of the word w in one of the two sets ($\mathcal{T}_c$ and $\mathcal{T}_{\overline{c}}$) and low frequency of occurrence in the another. The log part represents the popularity; more positive: terms are more popular in clicked set and more negative: terms are more popular in the not clicked set. We define the set of the

Table 1: Characteristics of the selected query sessions which expand more than 1 pages.

#query ses.	query length	% of query sessions which expands beyond page 1
568	3.44(Avg.)	19.71
94	1	67.02
251	2 to 3	47.8
140	4 to 5	56.86
83	>6	55.54

words interested to the user as $\omega = \{w|w \in \mathcal{W}, d(w) \geq \Theta\}$ and the set of words not interested to the user as $\overline{\omega} = \{w|w \in \mathcal{W}, d(w) \leq \overline{\Theta}\}$ where Θ and $\overline{\Theta}$ are the thresholds. We have considered $\Theta = 0.5$ and $\overline{\Theta} = -0.5$. The reason for ignoring the words $\{w|\overline{\Theta} < d(w) < \Theta\}$ is that they are likely to be noisy.

Given a new result $t \notin \mathcal{T}$, the weight of the user's interest on the result t is defined as follows.

$$f(t) = \sum_{w \in (\omega \cup \overline{\omega}) \cap t} d(w) \quad (2)$$

If $f(t)$ is positive, then we consider t as interested result to the user. In this paper, we focus on predicting the user's likely to be interested results.

3. EXPERIMENTAL VERIFICATION

All the experimental data used for discussion in this section are collected using a tool[1] which were used by a small community of around 12 users. This tool is a metasearch engine which monitors and records the users behaviour and performs few possible optimizations. For each new query request, the log contains the query, query session id, time of the request, ip address. For each click on the results, the log contains the url, rank of the url in the list of the results, time of the click, query, session id, ip address. It also maintains a cache containing detail information of the results and the clicks for every query session. It allows the tool to process the request in real time. Table 1 shows the characteristics of the query sessions. From the complete log data, we have considered only the query sessions which have expanded at least upto two pages. We assume that expanding to next page means, user is not satisfied with the results in the first page. The entries in the 2nd row and 3rd column clearly indicate that users are not satisfied with the results in the first page (at least) in 19.71% of the total query sessions. From the third row to sixth row of Table 1, it clearly shows that the query session with very short and very long queries are more likely to be expanded compared to the query sessions of query lengths 2 to 3 words.

We extract the unigrams and bigrams from the snippet of each result and $\mathcal{W}$ (see Section 2) is a set of the extracted unigrams and bigrams. Therefore, the user's interest is defined by the set of unigrams and bigrams. We ignore stopwords.

We predict users interest ω from the clicks and not clicks in the first page. Using the unigrams and bigrams thus obtained using Equation 1, we predict the list of the results of users interest using Equation 2. These predicted lists from the first page are then compared with the actually clicked results by the user in the next page of the expansion. We expect that user should click on the results predicted by the system. We define the accuracy of our prediction of each query session s by the following ratio.

$$accu(s) = \frac{\#\text{actually clicked results out of the predicted list}}{\text{Total \# of clicks in the expanded page}}$$

Table 2: It shows the performance of the proposed mechanism in predicting user's interest in next page.

session with query length	Avg. Accuracy	#Avg. predicted results
1	0.97	0.46
2 to 3	0.96	0.57
4 to 5	0.99	0.78
>6	0.99	0.76

Table 2 shows the performance of our prediction. For long queries (4th and 5th row of Table 2), almost all the results in the next page are covered by the predicted list. So, it is obvious that most of the user's clicks (99% in Table 2) are among the predicted list because of its large coverage. On an average, there is not much significance in predicting user's interest for long queries.

But for short queries (2nd and 3rd row of Table 2), the predicted list covers only a small portion of the results in the next page(less than 60% on an average). Even with such small coverage, it predicts with very high accuracy.

4. CONCLUSION

In this paper, we attempt to identify users current interest using click information submitted by the user at the time of search and optimize the results in real time. From the experimental verification, it is found that the proposed mechanism can predict users interest with an accuracy of 96.5% in real time for short queries.

5. REFERENCES

[1] Sevada. `http://www.lantana.tenet.res.in/~ranbir/sevada`.

[2] N. Alemayehu. Analysis of performance variation using query expansion. *Journal of the American Society for Information Science and Technology*, 54(5):379–391, 2003.

[3] B. J. Jansen, D. L. Booth, and A. Spink. Determining the user intent of web search engine queries. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1149–1150, New York, NY, USA, 2007. ACM.

[4] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 391–400, New York, NY, USA, 2005. ACM.

[5] F. Qiu and J. Cho. Automatic identification of user interest for personalized search. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 727–736, New York, NY, USA, 2006. ACM.

[6] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217, New York, NY, USA, 2004. ACM.