The Optimal Stopping Method for Search Engine Returning Trustworthy Results

Huang Jing-jing, Zeng Guo-sun

Department of Computer Science and Technology, Tongji University, Shanghai 200092, China
111huangjj@tongji.edu.cn, gszeng@tongji.edu.cn

Abstracts - The search engine generally returns a large number of results corresponding to a given query, making it more difficult for users to get the information they really look for. This paper gives a research on the optimal stopping method for search engine returning trustworthy results: firstly, it describes the stopping problem in search engine, then combines the candidate pages' trust degree, which is evaluated by Trust Facts Based Trust Evaluation Method, with their searched cost to get the searched reward, then gives a limited condition to determine how to select trustworthy results .Secondly, the conclusion of optimal stopping is applied to design an optimal selection method. Finally, we implement this method, and the experiment shows that our method gathers more correlative and trustworthy results for user's query, and achieves preferable effects.

Index Terms - Search Engine, Stopping Problem, Trust Degree, Searched Reward, Optimal Stopping.

I . Introduction

With the rapid spread of web information, it's more difficult for users to get the information they really look for. And one of related studies in [1] shows that almost 45% of all users click the first rank page, while 13% click the second one, only 3% will look the tenth page. The search engine usually returns numerous and redundant results for a user's query.

A lot of related work about search engine optimization has been done by researchers. Reference [2] proposed a personal search engine model based on users' behaviours and modified PageRank method in 2009. This model recorded user' queries and the corresponding pages users clicked, then it calculated a new page rank value for all results based on user interest and the PageRank, thus returning more relevant and acceptable results without asking the user about any information. In 2011, Reference [3] introduced a linear programming mathematical model for optimizing the PageRank of results. This method assigned a natural weight to each result of a group of web search engines for a given query, and then applied linear programming to get a new page rank for each result. As how to save users' reviewing time, Reference [4] studied users' behaviours and proposed an optimal strategy for reviewing search results. This method modelled the task of reviewing search results using two rational agents to maximize the probability of finding a satisfying result by giving some cost to reviewing results.

All of these methods optimize the PageRank of results or save user's browsing time, but the search engine still returns numerous results, even the false and irrelevant information. This paper will study how to select trustworthy results to improve the users' searching experience and satisfaction and search engine's efficiency.

II . Stopping problems in search engine field

A. The Stopping Problem in Search Engine

Stopping Problem is a common problem in real life, such as Classical Secretary Problem in [5]. As how to choose the best one, every selected object could be seen as a random variable, and choosing which variable depends on the conditions of objective function.

Def.1 Stopping Problem: Observing a sequence of random variables $x_1, x_2\ldots$ as long as you wish. For each $n = 1, 2...$, after observing $x_1, x_2, \ldots, x_n$, you may stop and receive the known reward $y_n(x_1, \ldots, x_n)$ (possibly negative), or you may continue and observe x_{n+1} . If you choose not to take any observations, you receive the constant amount y_0. If you never stop, you receive $y_\infty(x_1, x_2,\ldots)$. The stopping problem is to choose a time to stop to maximize the reward or maximize the probability of choosing the best reward.

Def.2 Search Engine Stopping Problem: Considering that, there a given keywords, the search engine has a set of candidate pages $R= \{page_1, page_2, \ldots.,page_n, \ldots|n\in N\}$. The search engine may keep observing results and outputting them to user. There exists a problem: when the search engine could stop to make sure all of the results are enough for user, and all of them are relevant to user's query.

If the search engine stops early, then the results are insufficient for user's query, but if the search engine outputs every page it has for user, even the fraud information, it will cost excessive resources, such as storage, bandwidth, and time. The motivation of this paper is to make the search engine select the high trust degree results and output proper number of results for users.

B. The Solution of Optimal Stopping Problem

In the general solution of stopping problem in [5], there are a sequence of random variables $x_1, x_2, \ldots,x_n, n\in N$, and the real-value rewards $y_1(x_1), y_2(x_1,x_2),\ldots, y_n(x_1, \ldots, x_n)$. Now we reject the first r-1 variables and then accept the next relatively best variable at the r stage, if any. The probability of win using this method is

$$P_r = \sum_{k=r}^{n} P(k^{\text{th}} \text{ variable is best and is selected})$$

$$= \sum_{k=r}^{n} P(\text{best of first } k-1 \text{ appears before stage } r) * \frac{1}{n}$$

$$= \sum_{k=r}^{n} \frac{1}{n} * \frac{r-1}{k-1} = \frac{r-1}{n} \sum_{k=r}^{n} \frac{1}{k-1}$$

The optimal r is the value that maximizes P_r, since

$$P_{r+1} \le P_r, \quad \text{if and only if}$$

$$\frac{r}{n} \sum_{r+1}^{n} \frac{1}{k-1} \le \frac{r-1}{n} \sum_{r}^{n} \frac{1}{k-1}, \text{ if and only if, } \sum_{r+1}^{n} \frac{1}{k-1} \le 1$$

Then, $r = \min\left\{ r \ge 1 \text{ and } \sum_{r+1}^{n} \frac{1}{k-1} \le 1 \right\}$

Hence, for large n, it is approximately optimal to pass up a proportion P_r =1/e ≈0.368, that's 36.8% of the random variables and the select the next relatively best variable as the final choice. As a conclusion, variables in the 36.8% are relatively good of the whole variables.

III. The Optimal Stopping Method of Search Engine Returning Trustworthy Results

As is known to all, the information text of a page consist of sever several sentences, and the trust degree of the text could be measured by the trust degree of the trustworthy sentences, which we define as Trust Fact. In this paper, the Trust Fact Based Trust Evaluation Method of Information Text in [6], is used to evaluate the trust degree of web pages and directly support for optimal stopping method to filter false results.

A. The Trust Degree of the Trust Fact

Declarative sentence is the expression of relatively complete linguistic unit, it is a statement of a fact or a point, and it also has the statement type which most embodies the text message content. The Trust Fact Based Trust Evaluation Method makes use of declarative sentences in the text of a page to calculate the trust degree of this page.

Trust fact (*TF*) is a declarative sentence which describes the concept and attributes of a special object, denoted by *f* in this paper.

We define *TF* like this: *f* = (*TO*, *TM*, *TP*, *TC*), and *TO* (Trust Object) is the object the *TF* describes; *TM* (Trust Magnitude) is the magnitude of the *TF* judgment predicate; *TP* (Trust Predicate) means judgment predicate, it has both positive form (f^P) and the negative one (f^N); *TD* (Trust Description) is the description of *TO* in *TF*.

According to the different forms and magnitudes of judgment predicate, a *TF* could be six forms of two groups, that is positive group: $f^{PG}=\{f_{TC}^P, f_{TD}^P, f_{TP}^P\}$, $f_{TC}^P, f_{TD}^P, f_{TP}^P$ respectively represent complete affirmative *TF* , general affirmative *TF* , part affirmative *TF*; the negative group: $f^{NG}=\{f_{TC}^N, f_{TD}^N, f_{TP}^N\}$, $f_{TC}^N, f_{TD}^N, f_{TP}^N$ are complete negative *TF* , general negative *TF* , part negative *TF*.

The Internet is a huge repository of information, and we use the six forms of a *TF* as keywords to get the number of results by keywords based full-text retrieval technique. If the search engine supports this *TF*, it will return more exact math results and make this *TF* a high proportion value. Therefore, the trust degree of *TF*, which is donated by *TD* (*f*), is evaluated as follows:

$$TD(f) = \frac{\sum_{s \in f^{PG}} N(s) * G(s) / G(f)}{\sum_{s \in f^{PG}} N(s) * G(s) / G(f) + \sum_{s \in f^{NG}} N(s) * G(s) / G(f)} \quad (1)$$

, where *s* is one of the six forms of *TF*, that is, $s \in \{f_{RC}^P, f_{RD}^P, f_{RP}^P, f_{RC}^N, f_{RD}^N, f_{RP}^N\}$, $G(s)/G(f)$ is the support between *s* and *f*.

In order to make sure whether the *TF* is related to given keywords, we introduce the correlation factor, that is, the keywords coverage of *TF*, donated by *C*(*f*), where $C(f) \in \{C_{all}, C_{part}, C_{none}\}$. Then the modified *TD* calculation can be defined as

$$TD^*(f) = \begin{cases} TD(f) * C_{all} \\ TD(f) * C_{part} \\ TD(f) * C_{none} \end{cases} \quad (2)$$

where C_{all}, C_{part}, C_{none} are all keywords coverage, partial keywords coverage, none keywords coverage in *TF* , we could give different weight, such as 1,0.8,0.5 respectively.

B. *The Trust Degree of Candidate p*ages

If $TD(page_i)$ is the trust degree of information text of candidate $page_i$ in *R*, then it could be described as the expectation of all TRs in this page, and could be calculated as (3):

$$TD(page_i) = \frac{\sum_{f \in FS_T} TD^*(f)}{|FS_T|} \quad (3)$$

where *TD*(*f*) is evaluated by (3); FS_T defines the set of *TF* in $page_i$, that is $FS_T = \{f_1, f_2, \ldots, fn\}$, and $|FS_T|$ is the number of *TF* in $page_i$. However, some important factors which have a significant effect on *TD*, such as where *TF* appears, are not manifested in (3). For example, if the *TF* appears in the abstract of $page_i$, that means this *TF* has a more great influence on the *TD* calculation, and we should give this *TF* a bigger weight. According to this, equation (3) becomes

$$TD(page_i) = \frac{\sum_{f \in FS_T} TD^*(f) * W(f)}{|FS_T|} * \frac{1}{1 + \frac{1}{4} log(|T_S| / |FS_T|)} \quad (4)$$

where T_s is all the sentences in $page_i$, that is $T_s = \{s_1, s_2, \ldots, s_m\}$, and $|T_s|$ is the number of the sentences in T_s. And $W(f)$ is the weight of *TF*, it's expressed as $W(f) = \varphi + m * W_P(f) +$

$n*W_K(f) + q*W_R(f)$, the value of coefficient φ, m, n, q could respectively be 0.7, 0.1, 0.1, 0.1, what's more, W(f) depends on the following three factors: (1) $W_P(f)$, which is the position of *TF* in $page_i$. According to the position of *TF* in abstract, the head of paragraph, the end of paragraph, the middle of paragraph, we could set ratio of 5:4:4:2. (2) $W_K(f)$, which means the keywords frequency of *TO* of *TF* in $page_i$. (3) $W_R(f)$, which is the topic relevance between *TF* and user's query. Further interpretation and application are in [3].

C. The Searched Cost of Candidate Pages

If the search engine uncritically store or output the results, then it will pay a huge storage cost. Actually, time consuming, occupied bandwidth and energy consumption are all unavoidable when search engine grabs web pages. What's more, it costs a lot of bandwidth or mobile phone traffic when user click these pages. More important, if the information text of these pages is difficult to understand, it will take up user's energy and time to read the information. Therefore, searching and outputting results cost a lot, and in this paper, we mainly analysis following aspects:

(1) Storage Cost: The bigger candidate page size will cost more data space. And the storage is based on text size ($S_{text}(i)$), picture size ($S_{picture}(i)$), and video size ($S_{video}(i)$), after search engine preprocessing, and the corresponding unit prices are α, β, γ, Then Storage Cost is computed as in (5):

$$C_{SC}(page_i) = S_{text}(i)*a + S_{picture}(i)*\beta + S_{video}(i)*\chi \qquad (5)$$

(2) Network Traffic Cost: The network traffic is mainly from two aspects, one is from search engine web crawler ($T_{crawler}(i)$), and another is generated when user click this page, donated by $T_{user}(i)$. Assume the unit price of network traffic is σ, the Network Traffic Cost is as follows:

$$C_{NTC}(page_i) = \left(T_{crawler}(i) + T_{user}(i)\right)*\sigma \qquad (6)$$

(3) Text Comprehending Cost: When the candidate pages' information is untrusted or irrelevant, user needs to spend more effort to understand what he is reading. Our algorithm for Text Comprehending Cost depends on the number of sentences which respectively cover all keywords ($S_{all}(i)$), partial keywords ($S_{part}(i)$), none of any keywords ($S_{none}(i)$) in $page_i$:

$$C_{TCC}(page_i) = \lambda * \frac{S_{none}(i)}{S_{sum}(i)} + \mu * \frac{S_{part}(i)}{S_{sum}(i)} + \nu * \frac{S_{all}(i)}{S_{sum}(i)} \qquad (7)$$

where $S_{sum}=S_{all}+S_{part}+S_{none}$, λ, μ, ν are the corresponding unit prices.

(4) Reading Time Cost: The easier the information in $page_i$ is, the fewer time the user will spend, $T_{time}(i)$ is the time user finish reading the text information of $page_i$. And the Reading Time Cost is computed as (8):

$$C_{RTC}(page_i) = T_{time}(i)*\rho \qquad (8)$$

where ρ represents unit reading price. In order to simplify the analysis, we consider these main four aspects. The searched cost (*SC*) for $page_i$ could be evaluated by (9):

$$C_{SC}(page_i) = C_{SC}(i) + C_{NTC}(i) + C_{TCC}(i) + C_{RTC}(i) \qquad (9)$$

Values of all the unit prices can be defined according to actual situation. For example, if we assume α=0.00300¥/KB, β=0.00040¥/KB, σ=0.0003¥/KB, ρ=0.0022¥/min, and apply this values into a report named "To unwaveringly uphold and develop socialism with Chinese characteristics" from People's Daily in [7]. There are 3 sentences covering all of the keywords, 19 covering partial keywords, 14 covering none of any keywords in this report. And we image user could read 600 words per minute, then the C_{SSC}= 0.0815¥, C_{NTC}=0.0072¥, C_{UBF}=0.4280¥, CR=0.5167¥. Finally, we evaluate the searched cost of this report is 0.5167¥.

D. The Search Engine Reward

If a candidate page has higher trust degree, has more relevant and easier understanding information, takes less time or energy consumption, it is really necessary for search engine to return such valuable result to users. But otherwise is not. Comprehensively think about both trust degree and searched cost, we introduce searched-reward, donated by $Reward(page_i)$ to evaluate the value of candidate pages as (10):

$$Reward(page_i) = \frac{TD(page_i)}{C_{SC}(page_i)} \qquad (10)$$

E. The Optimal Stopping Method

Searched-reward (*SR*) is a quantitative basis for search engine whether return the pages. Obviously, higher *SR* of one page means higher *TD* and lower *SC*, and results like this kind will improve user's searching satisfaction. Therefore, we give a description of Optimal Stopping Method of Search Engine Returning Trustworthy Results as follows:

Considering a given keyword, the search engine has a set of results R= {$page_1$, $page_2$, ….,$page_n$, …|$n\in$N}. The search engine keep observing pages in set R and calculating Searched Reward of them at the same time until the $page_{0.368|R|}$, where $|R|$ is the size of set R. Assuming θ_0 is practical threshold, if the Searched Reward of candidate pages satisfies the following condition:

$$Reward(page_i) \geq \theta_0, 1 \leq i \leq 0.368|R| \qquad (11)$$

Then search engine selects this page, otherwise pass it up.

F. The Optimal Stopping Method of Search Engine

According Definition 4 and the Optimal Stopping Method, we design the Optimal Stopping Method of Search Engine as follows:

Algrithm1:OptimalStoppoingMethodForSearchEngine()
Input：Keywords, Threshold θ_0
Output：Optimal results set R^*

R = {$page_1$,$page_2$,$page_3$,…,$page_n$}← *all_results*(*Keywords*);
R^1 = {$page_1$,$page_2$,$page_3$,…,$page_{0.368|R|}$};

for each page $page_i \in R^1$ do

{for all sentences $s \in page_i$,do

{ $f \leftarrow$find_sentence_fact (s); // find all TF in $page_i$

$F[f] \leftarrow format(f)$; // $F[f] = \{f_{RC}^{P}, f_{RD}^{P}, f_{RP}^{P}, f_{RC}^{N}, f_{RD}^{N}, f_{RP}^{N}\}$

$TD(f) \leftarrow Google\_search(F[f])$; // the TD of trust fact

$TD^*(f) \leftarrow TD\_of\_TF(f, C(f), N[f])$;

$W \leftarrow weight\_of\_fact(f)$; // the weight of trust degree

degree[ ]$\leftarrow TD_S$, weight[ ] $\leftarrow W$; }

$\boldsymbol{TD(page_i)} \leftarrow page\_trust\_degree(degree[], weight[], TD^*(f), |S|)$;

// Searched cost of candidate pages

{ $(S_{text}, S_{picture,}, S_{video}) \leftarrow Storage\_Space(page_i)$;

$(S_{none}, S_{part}, S_{all}) \leftarrow SentenNum\_based\_Key\_Cov(page_i)$;

$(T_{crawler}, T_{user}) \leftarrow Compute\_Traffic(page_i)$;

$T_{time} \leftarrow Reading\_Time(page_i)$;

$C_{SC}(page_i) \leftarrow Storage\_Cost([S_{text}, S_{picture,}, S_{video}], \alpha, \beta, \chi)$;

$C_{TCC}(page_i) \leftarrow Text\_Comp\_Cost([S_{none}, S_{part}, S_{all}], \lambda, \mu, \nu)$;

$C_{NTC}(page_i) \leftarrow Network\_Traffic\_Cost([T_{crawler}, T_{user}], \sigma)$

$C_{RTC}(page_i) \leftarrow Reading\_Time\_Cost(T_{time}, \rho)$

$\boldsymbol{C_{SC}(page_i)} \leftarrow SC(C_{SC}(page_i), C_{TCC}(page_i), C_{NTC}(page_i), C_{RTC}(page_i))$ }// Searched reward of candidate pages

Reward($page_i$) $\leftarrow TD(page_i) / C_{search\_cost}(page_i)$ // Optimal stopping selection if Reward(pagei) ≥ θ0 // select the results satisfy (11)

$R^* \leftarrow R^* \cup page_i$;} return R^*; //return final results set

IV. The Searching Experimental Analysis

Using "2012 China Internet industry economic development" as the keywords through Google search engine, we get 520 results. We calculate the trust degree, searched cost, searched reward of these results.

During this experiment, we evaluate the candidate 191 (520*0.368) search pages excluding outliers, which is too high or too low. Set 0.7108 as threshold value θ_0, and then select the remaining candidate pages by the rules: the results with the decision threshold value greater than the return value are selected, or lay down, and eventually we get back a result set R^*, $|R^*|$=98.

Then, we recognized those 191 results manually, shown in Table I. Accurate rate, and recalled rate and $F1$ value are three species common assessment parameter of performance, which means: as for a given category, x is defined as text number that divided into the correct class, y is the text number but divided into other class mistakenly, z is the error text number divided into this class, $p=x/(x+z)$ is accurate rate, $r=x/(x+y)$ is recalled rate. $F1$ index is a comprehensive evaluation of the precision and the recall rate, which is defined as $F1= 2pr/(p+r)$.

As shown in table 1, this algorithm has a good precise rate and recalled rate of selected class and rejected class, but also slightly insufficient. The main reason is large repeat results in the search engine, these repeat and searched reward less than threshold value of results will be filtered successfully by algorithm. However searched reward greater than threshold value and repeat of results, was divided to rejected class in artificial recognition, but are selected in this algorithm recognition, which reduces the accuracy.

TABLE I Performance Analysis of Optimal Stopping Method in Search Engine

	artificial recognition	Algorithm recognition		Precise	Recall	F1
		select	reject			
select	101	78	23	0.7959	0.7723	0.7840
reject	77	20	57	0.6129	0.7403	0.6707
unknown	13	0	13	0	0	*

V. Conclusion

This article for the problems of excessive results returned by search engine, proposes an optimal stopping method. As shown in our experiments, compared to the traditional methods of search engine optimization methods, ours can be more useful to help the search engine filter untrusted, irrelevant results, and return quality results to the user.

However, there are some shortcomings that return duplicate results with high searched reward. Compared to current search engine with a lot of irrelevant results, this method returns accurate results. Even if repetitive, the algorithm is of great reference value, therefore it is of real practical significance.

The shortcoming will be improved in the further work. At the same time, we will explore relevant factor assignment of different queries found to the topic, make the algorithm have a general meaning.

VI. Acknowledgment

This work was supported by the National Natural Science Foundation of China (grants 61103068, 61272107, 61202173), and the special Fund for Fast Sharing of Science Paper in Net Era by CSTD under grant No. 20110740001.

References

[1] T. Joachims, et al, "Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search," *ACM Transactions on Information Systems*, vol. 25, no. 2, pp.1-27, April 2007.

[2] M. Harb, R. Khalifa, and M. Ishkewy, "Personal search engine based on user interests and modified page rank," Proceedings of ICCES, IEEE Press, 2009, pp. 411-417.

[3] R. Amin, E. Ali, "Optimizing search engines results using linear programming," *Expert System with Applications*, vol. 38, no. 9, pp. 11534-11537, September 2011.

[4] J. Huang, A. Kazeykina, "Optimal Strategies for Reviewing Search Results," Proceedings of the National Conference on Artificial Intelligence, 2010, pp. 1321-1326.

[5] T. S. Ferguson, "Who solved the secretary problem," *Statistical Science*, vol. 4, no. 3, pp. 282-289, August 1989.

[6] D. Q. Zhang, G. S. Zeng, Wang Wei, "Trust facts based trust evaluation method of information text," *Computer Science*, Vol. 35, No. 8, pp. 202-205, August 2008.

[7] Z. J. Li, "Unwaveringly uphold and develop socialism with Chinese characteristics," http://paper.people.com.cn/rmrb/html/2013-01/06/nw.D110000renmrb_20130106_2-01.htm?div=-1