The Research on Personalized Query Expansion Technology

SongYi[1, a], Zhang Shi Long[2,b] and Wang Jia Ning[3,c]

[1, 2, 3] Dept. of Computer, HuaDe Universityof Harbin, China

[a] songyiyouxiang@163.com

Keywords: Personalized; Query; Eextension Technology.

Abstract. The current information retrieval system, there are still some problems, such as incremental information collection of web crawler, index dynamic updating and compression, results sorting, audio and video retrieval, query language visualization and so on. The results returned users are most concerned about whether to meet the needs of retrieval. Query by example "rocket", for sports fans user search intention is "rocket" related to the sports information, and of military enthusiasts search intention is indeed "rocket principle" or "rocket launch" military category information. Usually user input query string is shorter, search engine is very difficult to identify the user interests preference categories, in this paper inspired if the query expansion, with the user query related information into the query expansion base, feedback to the search system, add user interest model, identification of user interests to be more clear identification of user interests preference categories. In this paper, based on dictionary and based on relevant feedback combined with method of personalized query expansion and effectively identified user interests preference categories.

Introduction

At present most of the search engine is that the retrieval based on keyword matching. Due to the lack of understanding of semantic keywords this method, search results to the user is not ideal, traditional search query and matching basic hypothesis is users want the page contains the input text queries [29], but users do not pay much attention to page can contain the query itself, often included in the web page is user query related or the close synonyms. All users enter the same query condition, the search engine will return the same results, although the user's needs are not the same. Traditional retrieval systems do not consider the user's individual needs, the search results are not related to the web too much, in all the results are often the most of the results and the user's demand is not. Although some web pages containing the search key, with the intention of the user actually has nothing to do, but it is returned to the user.

Personalized user search model is a key issue, how to identify user preference category is the key, that is to build user interest model. Massive network search engine users to log the user behavior of deep mining information retrieval user behavior information by user submitted query, the user click and access page log information is acquired, and real user's search intention and demand is hidden behind these logs, cannot be obtained directly. Even get these indirect information also exists in a lot of noise, the noise can be by noise and reliability of network data itself, may also be due to the uncertainty of user behavior. So a user query expansion based on other content of technology, the deep mining of user behavior is the key and difficulty of the research should first be solved.

User Behavior Mining

According to the research of user's interest preference, the existing large-scale Internet user log mining is used to mine the user's search behavior. This work is based on a large-scale search engine log, including the integrity of the log query, click behavior, browse the contents of the link and the corresponding timing information, etc.. Based on the analysis of large scale log data, the common user behavior features are extracted, and then the user interest model is constructed. Specific methods, the use of the query word and the user clicks on the page of information can be analyzed the general characteristics of the user query. Query length, frequency, query based content, theme classification

and clustering analysis etc.; the browsing time and browser identifier to extract the characteristics of a specific period of time within the same user query and modify the query process, to establish the user query correction model and user query language model; using the user clicks on the results page and click on the information, such as query result ranking and user clicks, access and the length of time the computer IP address, users click patterns mining; users click on interest page and to the theme of the distribution characteristics of deep layer analysis, user behavior between the user query and click on the relationship between deep mining. The research work in this paper is mainly based on the macro analysis of the bulk of the user requirements, the main purpose is to find the user demand hot, word frequency distribution and query behavior characteristics such as, and the retrieval system system structure and algorithm design to make improvements. Basic data for the user log analysis includes data for two days in March 1, 2015 and March 2nd, as shown in table 1.

Table 1Query Data

date	***Query Record***	***Query Stringy***	***URL***
3.21	135076	244501	560091
3.22	1228608	65536	4361025

Query length analysis

The length of the query, that is, the number of query words in the query string that the user entered. The length of the query string is shown in Figure 1, where the X axis represents the length of the query word, the unit is a unit, and the Y axis indicates the number of the corresponding keywords.

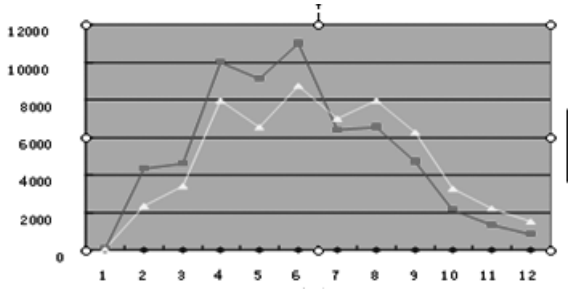

Fig.1 The length of the query

In the query log, the length of all the query string statistics, Figure 1 explicit query string length of 6 words of the most, followed by a length of 4 of the query string. It can be seen that the length of the distribution curve is fluctuating, and the length of the even number of bytes is significantly more than that of the adjacent odd number of bytes of query string. This is clearly and most query string containing Chinese character, a Chinese character takes two bytes. Enter a query string length statistics after the data information of integer numbers, ease of handling and analysis; URL meet heaps law so that users click on the URL was a high degree of repetition and explicit URL with local characteristics; users click on lean forward a few pages, sort of user queries is quite important.

Page number analysis

The results page to view the user input query request, statistics users to view the page number, such as flip, and other information, user page information table as shown in Figure 2.

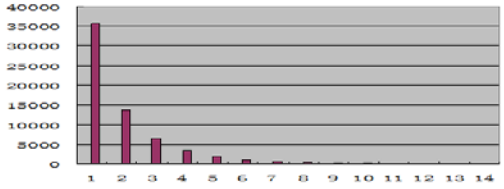

Fig.2 The page information in one day

The two day comparison page as shown in figure 3.

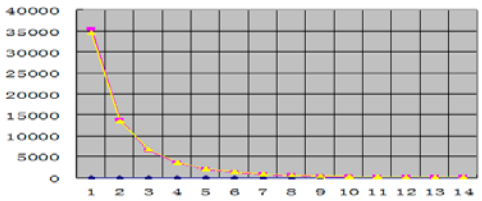

Fig.3 Comparation of page in two days

After taking the logarithm as shown in figure 4. In this paper, the user view the number of result pages statistics show that: about 54.24% of users see only the first result page, 21% query the first two pages of results, 10% of the users view the first three results page, only less than 0.42% of the users view the more than 10 results page. For the user to view the results page number is less and less, suggesting that by clicking on the URL with the local, users click on local inspired the search engine system as far as possible the results into the results of the first few pages explicit to the user, ensure the row in the first few pages query result is the high quality of query results, search engine ranking mechanism optimization.

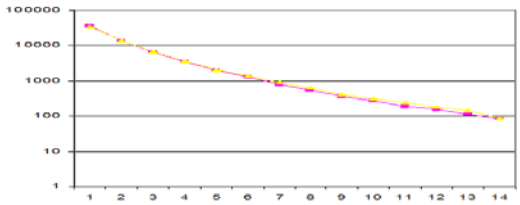

Fig.4 Comparation of page after logarithm

This paper can be found, user queries the average length is 2.1 query, query words short it is hard to show users to specific user interests and synonymy and polysemy phenomenon exists generally, this paper enlightens were query expansion.

Research on Query Expansion Technology

Every day, nearly one million users use the search engine to search relevant information to meet the needs of daily life. Query expansion QE (Expansion Query) refers to the original query based on the addition of the user input with the search terms associated with the new words. The new longer and more accurate [31] query, query expansion using [32] feedback technology. Early research by adding Related words is extended, commonly used query expansion method can be roughly divided into three categories: semantic knowledge dictionary method, global analysis and local analysis method based on. Semantic knowledge dictionary expansion method classification dictionary computing feature weights is to a single word as a unit, and each word may belong to multiple classes, lead to the most relevant category decision error, further affect the accuracy of retrieval results next. Local analysis when the initial query back in front of the documents related to the original query is not big, local analysis will be a large number of independent query expansion join query, which severely reduces the query precision, even lower than do expansion optimization of the case.

Personalized query expansion method

Personalized query expansion process

This paper found the user query is a cross class phenomenon, such as: the user submitted "rocket" to the search engine, then the user's query intention is "Sports" category, or belonging to the "Military" category? For example, users to search the "traffic" is to query the traffic situation, or bank credit card etc.. This paper separately according to the user's query is difficult to distinguish between user interest categories, so the personalized query expansion, through query expansion, the "rocket" two kinds of related queries {Dallas, San Antonio, Houston Rockets Houston, Rockets game, Rockets fans, fire arrow schedules, basketball, rockets preseason, the Rockets man, NBA}, {off in a rocket, space shuttle, rocket propellant, moon rocket, rocket, long march rocket rocket} and then extends out of the algorithm through experiments, query expansion before and after experimental comparison, this paper can find user interested in sports categories, users have indeed participated in the

experiment in the field of sports a high degree of interest, which fully shows the expansion after the effective improvement of the recognition problem of cross category query. Query expansion based on dictionary and query expansion based on relevance feedback. Expand query with extended thesaurus, feedback to the user interest model, to calculate the degree of user interest, the specific calculation method will be discussed in detail in chapter four. Personalized query expansion process.

Query Expansion Based on Dictionary

The extended dictionary method based on thesaurus dictionary as famous. Pinyin input method can cover almost all of the words in Chinese, so this paper dictionary use famous cell thesaurus, famous cell thesaurus 11016 entries, including eight categories, 49. For example, sports: football, basketball, fitness, track and field etc. Each category containing a thesaurus, such as basketball (23 entries), basketball star (718 entry), NBA team name (57 entries), basketball terms (228 entries), basketball glossary (2384 entry), NBA player name (75 entries), NBA (43 entries) and sports special (621 entry). Users enter the query first to scan the dictionary, in the dictionary for the longest matching search process, that is, enter the query sequence, find the sequence in the famous dictionary said there is the longest match. If there is a query string, and the word added to the extended lexicon. For example, the input query is extended to the traffic, traffic, traffic: banks and so on.

Query expansion based on relevance feedback

The expansion of the ideological source for relevance feedback model Rocchio algorithm, according to the related documents and relevant documents for query modification, and thought is the Related words in the relevant document distribution is relatively uniform, in the no relevant document is sparse, as shown in equation 1.

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d} \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d} \in D_{nr}} \vec{d}_j \qquad \text{Eq. 1}$$

qm———Modified query vector; Q0 represents the initial query vector;

α， β， γ———Adjust parameters, often 1;

Dr———Related document collection;

Dnr———Unrelated document collection。

Query expansion in the category of documents related to the query feature extraction, extraction of feature words. Then calculate the similarity query and the feature words, high similarity join query expansion. The query expansion characteristic is the I related documents is calculated by the model system, the user does not need to distinguish relevant documents, reduce the burden on users. As shown in equation 2

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|N_i|} \sum_{\vec{d} \in N_i} \vec{d}_j \qquad \text{Eq. 2}$$

q_0———initial query;

q_m———Extended Query.

Query expansion process, we found that there were some of the words, but joined the expansion of the vocabulary, which in terms of system is noise information, researchers proposed expansion of the noise to affect the query performance, when the query expansion reached 25 [33], decrease of retrieval accuracy, to add 20 expansion words were improved user model.

Where WQD represents the weight of the query in the document, COOC (Q, q ') on behalf of the query Q expansion Q' credibility, the specific calculation method as shown in the equation 3.

$$cooc(q, q') = \frac{nqq'}{nq} \qquad \text{Eq. 3}$$

cooc(q，q’) ———query expansion into the query expansion Q 'Q' extension credibility;

nqq’ ———Query Q and Q 'and the number of times in the document;

nq———Query the number of times Q appears in the document.

Experimental content

The text corpus using a well-known site news corpus, the original query 320 query extended 920 query, query "rocket", for example, expansion after expansion in the military and sports categories such as table 2 shows.

Table 2 Query Extend libarary

sprot		***militriry***	
rocket	Calf	rocket	blast off
Spurs	Houston	shuttle	rocket
rocket	match	rocket	propellant
fan	Basketball	moon	rocket
rocket	Schedule	carry	rocket
Pre season	Celts	march	rocket
Lakers	NBA	Space	launch

Query classification accuracy

The precision is widely used in the field of information retrieval (precision) and recall (recall) to evaluate the experimental results. The precision and recall is defined as shown in equation 4.

$$P = \frac{Q_T}{Q_A} \qquad \text{Eq. 4}$$

Q_T ———Correct number of queries;

Q_A ———Number of queries.

Query string has the corresponding class, the essence of this model is to query classification, query classification accuracy rate to evaluate the accuracy of classification. Enter the query string 320, respectively, belong to sports, military, automotive, education, tourism, IT six categories, the average accuracy rate of 0.86, the classification performance of each category as shown in Table 3.

Table 3 accuracy of query classification

sport	***military***	***education***	***automobile***	***tourism***	***IT***	***average***
0.92	0.87	0.88	0.81	0.82	0.89	0.86

Summary

In this paper, we first introduce the research of personalized search technology and key technology of personalized search, search and personalized user's interest and preference learning acquisition method, also in a large-scale search log the query length and page information analysis and experiment results are given. Introduce a query expansion technique for personalized query expansion, through query expansion to form an extended lexicon, the dictionary and improved Rocchio relevance feedback combined with query based on extension method for query expansion. Through query expansion technology, solve the user query string is short, the user query ambiguity. At the same time, the query expansion technique is applied in the user interest model can effectively identify the user class belongs to the category of query, such as user input "rocket", we do not know interested users to sports category "rocket", or categories of military "rocket" interested. But through the query expansion technology, the sports and military categories related to the expansion of the query information, to clearly identify the intention of the user query. Therefore, the query expansion technique is a good foundation for the recognition of user interest model interest lay.

References

[1] Setphen Akuma,Rahat Iabal,Chrisina layne.Comparative analysis of relevance feedback methods based on two. omputers in Human Behavior.(2016),p.138-145

[2] Lai Jiang, Runming Yao.Modelling personal thermal sensations using C-Support Vector Classification (C-SVC) algorithm. Building and Environment,USA, (2016),p.98~106

[3] XueHua, ShenBinTan, ChengXiangZhai. Implicit User Modeling for Persona-lized Search. (2016),p.5~6

[4] Salila Vongkrahchang, Apasara Chinwonno.Effects of Personal Intelligence Reading Instruction on personal intelligence profiles of Thai university,. Kasetsart Journal of Social Sciences. (2016),p.7~14

[5] Elena S. Kiseleva, Oksana V. Anikina.Modern Model of Competences of Personal Agents as Increase Factor of Clients' Subjective Well-beingOriginal. Proceedings of the 28th Annual International Procedia - Social and Behavioral Sciences, (2016),p. 116~121

[6] Hamid Hassanpour, Farzaneh Zahmatkesh. An adaptive meta-search engine considering the user's field of interest Manage-ment, Journal of King Saud University - Computer and Information Sciences, (2016),p. 71~81

[7] Gabriella Pasi.Inferring Implicit Feedback through User-system Interactions for Defining User Models in Personalized Search, Procedia Computer Science, (2016),p. 8~11

[8] Jaime Teevan, Susan T.Dumais, Eric Horvitz.Personalizing Search Via Auto-mated Analysis of Interests and Activities. Proceedings of the 28th An-nual International ACM SIGIR Conference on Research and Development in Infor-mation Retrieval, Salvador, Brazil, (2016),p.ACM: 449~451

[9] Feng Qiu, Junghoo Cho. Automatic Identification of User Interest For Perso-nalized Search.International World Wide Web Conference Committee, Edinbufh, Scotland, (2016),p. ACM: 23~26

[10] Paul Alexandru, Chirita, Wolfgang. Using ODP Metadata to Personalize Search. Proceedings of the 28th Annual International ACM SIGIR Conference, Aalvador, Brazil, (2005),p. 3~4

[11] Fan Liu, Clement Yu, Weiyi Meng. Personalized Web Search by Mapping User Queries to Categories. Conference on Information and Knowledge Management , Mclean, Virginia, USA, 2002. ACM: 5~6

[12] Yi Zhu, Li He, Xiaojun Wang. User Interest Modeling and Self-Adaptive Update Using Relevance Feedback Technology. Procedia Engineering, (2012),p. 721~725

[13] Maria Papadogiorgaki,Vasileios Papastathis,Evangelia Nidelkou, Simon Wad-dington,Ben Bratu, Myriam Ribiere, Ioannis Kompatsiaris. Two-Level Auto-matic Adaptation of a Distributed User Profile for Personalized News Content Delivery. Proceedings of the 28th Annual International Conference, Tampere, Finland, 2008. ACM: 1~3

[14] A. Eckhardt, T. Horváth , P. Vojtáš. A User Profile Learning Approach for Web Search. Intelligence.International Conference on Web Intelligence, Virginia, USA, 2007. ACM: 780~781

[15] ZhongMing Ma, GautamPant, Olivla , R.Liu Sheng. Interest-Based Personalized Search. ACM Transactions on Information Systems, New York, 2007. ACM: 1~5

[16] Xujuan Zhou, Sheng-Tang Wu, Yuefeng Li, Yue Xu, Raymond Y.K. Lau, Peter D. Bruza. Utilizing Search Intent in Topic Ontology-based User Profile for Web Mining. Proceedings of the 2006 IEEE International Conference on Web Intelligence, 2006. IEEE:1~3