Personalized Intelligent Search Engine Based on Web Data Mining

Hong Zhang¹, Yanhong Ma², Qiuyu Zhang¹, Pengshou Xie¹, Zhongxian Bao¹
1. College of Computer and communication, Lanzhou University of Technology, Lanzhou 730050, P.R.China.
2. Gansu Electric Power Corporation Wind Power Technology Center, Lanzhou 730050, P.R.China.

Abstract—A personalized intelligent search engine based on web data mining, whose theory model is established, is introduced and discussed and in this paper. And its critical part, algorithm and program of mining users’ interests, is described detailed. By creating users’ interest database, this system will realize personalized information retrieval which can search information according to users requests. So the recall and precision ratio of search engine is enhanced essentially.

Index terms—data mining, personalize, users interests, users pattern, search engine

I. INTRODUCTION

The study for search engine has always been a hot, which is due to the fact that there are many limitations of search engine [1]. Here are some examples.

i. Search engine can not be a good “understanding” of what users want to search, and only can match the key words or the sentence that user has input mechanically[2].

ii. Search engine does not have the nature of personalization [3]. No matter who is the retriever, research workers, businessmen, students, doctors and so on? As long as the key words input are same, the result returned is same.

iii. Search engine does not have the interactive nature. According to the returned results, users want to express their own wishes, but they could not do so[4].

In order to overcome these shortcomings of traditional search engines, an intelligent, especial tool for retrieving Internet information is needed to been developed urgently to help users get the information they needed from the Internet fleetly. On this basis, an intelligent, personalized search engine based on Web data mining is proposed[5], which is used to mine user's web history and track user's web acting by web data mining to create users’ interest pattern database, in which each user’s interest and hobby is stored. The interest pattern database is used to filter the user’s initial query results [6]. So the available information which meets users’ needs is returned to them and system realizes the personalized information retrieval.

II. THEORETICAL MODEL OF PERSONALIZED INTELLIGENT SEARCH ENGINE BASED ON WEB DATA MINING [7]

A. Gathering Information

Usually using an initial URL address as a starting point and utilizing the standard transport protocol, Robot, that is Web Spider or Crawler, ransacks all over the WWW space including all hyperlinks in Webpage to gather Webpage information and stores the information into Webpage database[8]. Compared with other Robot, this Robot can discover the dead links and find the newly added links by scanning web space constantly. There are two ways of obtaining initial URL address. One is collection by itself regularly and the other is referral by web site.

B. Processing And Indexing Information

The goal of this step is to extract available information from web pages which are gathered by Robot and index the information and build Index Database of web pages which can be retrieved by users. In some ways, it is the Index Database that determines the quality of search engine. So the design of Index Database is very pivotal. In order to incarnate this tenet, two measures are introduced in this system: knowledge database technique and new weighting algorithm which is introduced in references [9]

C. Retrieving And Matching Information

When users retrieve information, their interests or hobbies are taken into account [10]. It has high degree of specificity, not just to match users’ input simply. According to their query input and their interests or hobbies, it construct users search vector [11]. Namely, it refers to the users interests to determine whether the information retrieved is precise or satisfied with users. In addition, it can achieve active information push by judging user's interest, which is somewhat like the relevant information Push of web site Baidu. But there is essential difference between them. Because compared with information push of web site Baidu, this information push pays more attention to users’ interests. It is not only
the information push similar with the user's retrieval input.

D. Creating Users Interests Database

As we all know, if system want to achieve the personalized information retrieval, the first task is to know what the user's interest is. So system must has the function of storing user's interests, the function of feedback user's interests, the function of reasoning and judging user's interests and etc. [12].

User's pattern is used to resolve this problem. This module is mainly to integrate client user's web information and feedback information to mine user's web log to obtain users' interests or hobbies by using special web data mining algorithm which will be presented in the following text. And then a users interests database is created. As user's interest may change sometimes, the users interests database must be updated according to user's web record. It resolves such questions as: first, information received is difficult to be understood or is not precise. Secondly, Users don’t know how to express their requirements for internet resources appropriately or how to find the information they need effectively.

E. User Interface

User interface adopts browser, such as internet explore, to exchange data between users and servers. Users input query requirement, initial information and feedback information at client. The results are returned back to users by the form of browser too. By downloading Java Applet, Client communicates with server to achieve users' feedback and results transferring. Users can evaluate the retrieval results, such as best, better, good, no good and so on. These evaluations are feed back to the system to adjust user's interest information. So user interests are updated ceaselessly and always kept up to date. On retrieval interface, users can express themselves interests and correct, renew their interests database. Personalized intelligent Search Engine based on web data mining is designed to realize personalized information retrieval which can resolve the question that when different users use the same query, the results are different and that when a user uses the same query in different times, the results are different too.

Thus, personalized intelligent Search Engine based on web data mining is a typical case of using web data mining technology into personalized intelligent search engine. In the system, the most critical problem is to create user's interest pattern database by the way of using data mining techniques. Once user’s interest pattern database is created, system can combine user's interest into his retrieval input to provide him with more accurate and personalized results retrieved. Therefore, this paper gives the detailed process of using web data mining techniques to create users’ interests pattern database.

III. CREATING USERS’ INTERESTS PATTERN DATABASE

Each web server will keep the user's access information to it. Usually, this information is called WEB Log including web server access log, proxy server log records, Browser log records, Users’ brief introduction, users’ registration information and users’ dialogue or transaction information and so on[13]. The target of web data mining is to find the user's access pattern from vast amounts of web log data and to dig out available users’ information finally.

In order to obtain users’ pattern information and have real-time update for this information, system takes two steps to complete it: establishing users interests model and mining users interests.

A. Establishing User Interests Model

As we all know the real intent of this system is to achieve personalized information retrieval. So a data model must be created to do it. In this paper, users interest model is expressed by an ordered triad which is interested word, word weight, word fresh degree. Each interested node is marked with a triad \((p_i, w_i, x_i)\) abbreviated Node \((p_i)\) [14].
In above expression, the value range of \( p_i \) is \( P \), marked with \( p_i \in P \), and \( P \) is words sets, marked with \( P= \{ p_1, p_2, \ldots, p_m \} \), in which \( p_1, p_2, \ldots, p_m \) are the interested words and \( m \) is the number of words. The \( w_i \) is the weight of interested word \( p_i \); the \( x_i \) is the fresh degree of word \( p_i \).

For the sake of the fact that different location of word in the document reflects different importance, the location word appears is taken into account, which is called location weight marked with sign \( f(x_i) \). When calculating fresh degree of words, we use a fresh degree function \( f(n) \) to document \( d_n \) (\( d_n \in D \). Sign \( n \) refers to the nth document in buffers. Sign \( D \) is the document collection in buffers ). The function \( f(n) \) is monotonic and non-decreasing which can assure that the more recent a document is visited, the more users are interested in it. So the weight and fresh degree of Node \( (p_i) \) are calculated as follows.

\[
\text{Node}(p_i) \cdot \omega_i = \sum_{j=1}^{n} t_{f,j}^{w} \times E_j \\
\text{Node}(p_i) \cdot \chi_i = \sum_{j=1}^{n} \frac{t_{f,j}^{w} \times E}{\text{Node}(p_i) \cdot \omega_j} \times f(j)
\]

In above formula, the sign \( t_{f,j}^{w} \), \( p_i \), \( w_i \), \( x_i \), \( f(n) \), and \( n \) are explained as above. Sign \( E_j \) (\( E_j \in [0, 1] \)) is interest coefficient of document \( d_j \). And \( f(n) \) can be calculated by formula \( f(n) = \frac{n}{n+1} \). After the weight and fresh degree of word \( p_i \) is calculated, formula \( \chi_i = w_i \times f(x_i) \) is adopted to calculate interest degree of word \( p_i \). And \( f(x_i) \) is an influence function on fresh degree upon weight of word \( p_i \). It is calculated by formula \( f(x_i) = x_i \). Finally this information is stored into users interests database in the model of ordered pair which is expressed with the pair of interest words and interest degree. The interest degree of words is the ultimate basis for making search engine intelligent and personalized.

B. Mining Users Interest Pattern

Here are several ways of data mining is used to fond users interests.

a. Mining Association Rule [15]

Through correlation analysis, such as algorithm Aprior, relationships hidden among data are uncovered. Here are some examples. When mining association rules on web site server logs, we find that 70% users have accessed the football pages and 15% users have accessed the diving pages among users who have accessed sports news pages. Then such a conclusion can be drawn about: If a user likes sports, we can prediction that the probability he likes football is 0.70 and the probability he likes diving is 0.15. So if his query words contain the word sport, system will push the football pages to him but filter the diving pages. Then system will rectify the user’s interest parameter: the interest degree for sport is 0.7 and the interest degree for diving is 0.15.

b. Classification Analysis [16]

In the web log mining, the input set of classification analysis is group of record collection and several types of tags. First, each record is given a type tag. Then system checks these tags and describes the common features of these tags. For an example, 50% users live in large cities and their ages are between 18 and 28 among users who have submitted mp4 Orders. After getting this information, we can provide pertinent and personalized service to the aged between 18 and 28 users living in large cities.

c. Clustering Analysis [17]

Clustering analysis is different from classification analysis. It is the process of classifying data items or users with similar characteristics. For an example that some users often browse the pages about “TOFEI” or “GRF” or “application” or “visa”, then these users will be clustered as a group: they may be a group of expecting overseas users. Therefore system will send e-mail about going abroad to them and provide personalized service to them.

d. Sequential Pattern

Sequential pattern refers to find data items which are sequential in time from the time-series data sets. In the web log mining, sequential pattern recognition means to find the user’s requests for pages which are successive in time among user session. For an example that if 60% users ordering baby sleeping bag on line order baby clothes within 2 months, then system will predict the web pages that may be requested by the users and provide the users ordering baby sleeping bags web pages about baby clothes actively.

Of course, the methods of data mining are various. Only some methods on web mining are introduced here. When the user interest patterns are recognized, these patterns must be expressed with the formal language and formed knowledge or rules and stored in knowledge database which can be used when users retrieve.

IV. CONCLUSIONS

This paper analyzes the developing status of search engine and puts forward a new kind of search engine based on web data mining and gives its theoretical model. Each part of personalized intelligent search engine based on web data mining is described in detail. Several methods of web data mining in personalized intelligent search engine are introduced emphatically. By creating users’ interests pattern database, system combines user’s interests into his retrieval to achieve personalized information retrieval and information push service. So
users can get the exact information they want fleetly. The recall and precision ratio of search engine is improved.

As we know that personalized intelligent search engine based on web data mining involves not only data mining techniques but also artificial intelligence, pattern recognition, natural language retrieval, formal description and other related disciplines [18]. So it depends on the development of these related disciplines to achieve real personalized information retrieval.

REFERENCES


