Study of Distributed Personalized Search Engine

ZHANG Hong[1], MA Yan-hong[2] , MA Wei-jun[1], BAO Zhong-xian[1]
[1]College of Computer and communication, Lanzhou University of Technology,Lanzhou730050, P.R.China.
zhanghong@lut.cn
[2]Gansu Electric Power Corporation Wind Power Technology Center, Lanzhou 730050, P.R.China.
Myh_2005@163.com

***Abstract—Combining d*istributed computing and data mining techniques, a distributed personalized search engine is put forward to solve the the problems current search engines faced. It has described the theoretical model and algorithmic processing. Under the Hadoop, a distributed platform processing information with Java, the key parts are programmed and implemented. The experimental results show that this theoretical model can improve the accuracy and speed of the user's queries so it can improve the retrieval performance of the search engine.**

Keywords-distributed computing; personalized;search engine; Hadoop; user interests

I. INTRODUCTION

The rapid development and widespread popularity of the Internet has led to the updating of search engine and the emergence of distributed computing technology has injected search engine new vitality[1]. So it is irresistible to use distributed computing and personalized retrieval techniques to research distributed and personalized search engine architecture[2].

The earlier famous distributed search engine are Google, AltaVista's Internet Archive Crawler and Mercater, but they are deployed on the LAN and can only get smaller part information of Internet. In recent years, WAN distributed search engine are yacy and Faroo, which are relatively small in scale. In china, the research of distributed search engine is relatively few and the representatives are Skynet LAN-based developed by Peking University and IglooG grid-based developed by Shanghai Jiaotong university, which are all In the development stage[3]. no matter what type of search engines, they can not understand what users want to search, and only can match the key words or the sentence that user has input mechanically. Search engine does not have the nature of personalization. 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. 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.

In order to overcome these shortcomings of traditional search engines, a distributed personalized search engine has been proposed, 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[4]. So the available information which meets users' needs is returned to them and system realizes the personalized information retrieval.

II. THEORETICAL MODEL OF DISTRIBUTED PERSONALIZED SEARCH ENGINE

Theoretical model of distributed personalized Search Engine is shown in figure 1[5]. From Fig. 1 we can see it contains five components, which are distributed information gathering, distributed information indexing, distributed information retrieving, creating user interests database and user interface. Each part is described as follows.

A. Distributed Information Gathering

Usually using an initial URL address as a starting point and utilizing the standard transport protocol, each distributed Crawler crawls its WWW space to gather Webpage information and store the information into Webpage database. It also analyzes and extracts web links and produces a list of Web links library for the next crawling use. These Crawlers are distributed on each slave node in the master / slave structure and scheduled by scheduler on the master node[6].

B. Distributed Information Indexing

The goal of this module is to extract available information from web pages which are gathered by distributed crawler and index the information, calculate the webpage PageRank value, establish the inverted document library of webpages to complete the index database of webpages. Finally it realizes distributed storage of webpage files under Hadoop platform[7].

C. Distributed Information Retrieving

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.

Published by Atlantis Press, Paris, France.

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[8]. So system must has the function of storing user's interests, the function of feed backing user's interests, the function of reasoning and judging user's interests and etc.

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. Distributed personalized Search Engine 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, distributed personalized Search Engine is a typical case of using web data mining and distributed computing techniques. 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 user interests pattern database.

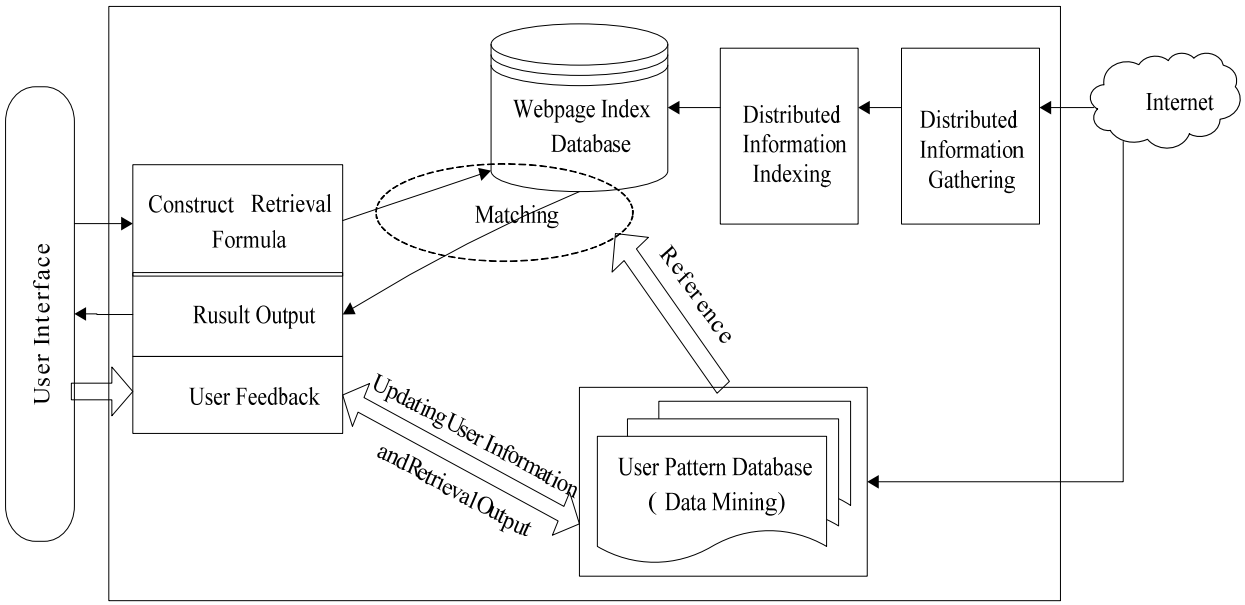

Figure 1. Theoretical model of distributed personalized search engine

III. ALGORITHMIC REALIZE OF DISTRIBUTED PERSONALIZED SEARCH ENGINE

Each web server keeps the user 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. 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.

So in the follow text, we interpret how to obtain and update users pattern information and how to implement distributed information retrieval.

A. User Interests Model Establishing

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) [9].

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、…、p_m}，in which p_1、p_2、…、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 $tf_{i,j}^{w}$ [9]. 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 monotonous 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.

$$Node(p_i) \bullet \omega_i = \sum_{j=1}^{n} tf_{i,j}^{w} \times E_j \qquad (1)$$

$$Node\ (p_i) \bullet \chi_i = \sum_{j=1}^{n} \frac{tf_{i,j}^{w} \times E}{Node\ (p_i) \bullet \omega_i} \times \mathrm{f}(j) \qquad (2)$$

In above formula, the sign $tf_{i,j}^{w}$, 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 $\mathrm{f}(n) = n/(n+1)$. After the weight and fresh degree of word p_i is calculated, formula $t_i = w_i \times f(x_i)$ is adopted to calculate interest degree of word pi. And f (xi) 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 [10] [11].

B. Distributed Information Retrieval

The implement of distributed information gathering and indexing is referenced literature[6] [7] which adopts Map/Reduce model to design and perform under Hadoop platform. Therefor, here the personalized information retrieval process adopting distributed computing techniques is expounded.

Distributed Information Retrieval is mainly responsible for responding user queries, retrieving information from Web information indexing libraries, after sorting out the information returning the results to users. It also adopts Map / Reduce model to program. system uses Tomcat6.0 as a Web server, Jsp/Servlet technology as programming environment. Its performing process is shown in figure 2.

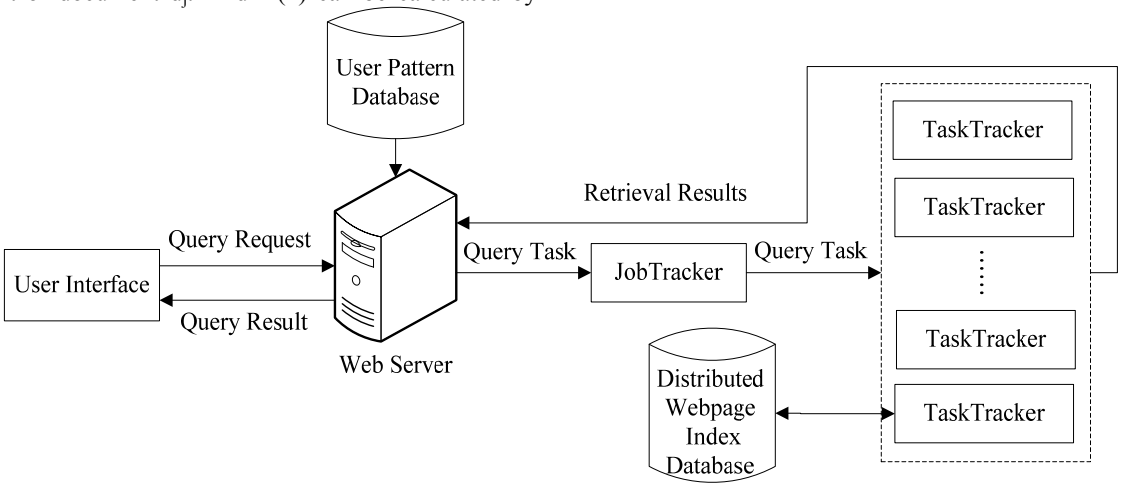

Figure 2. flow of distributed information retrieving

First users input query requests through user interface. After obtaining query requests, Web server analyzes and pretreatments these requests, extracts keywords, creates a query task combined with user interests and the task is send to each TaskTracker by JobTracker[12]. Then based on keywords, Map makes retrievals from distributed information indexing library and obtains indexing data. According to keywords Reduce summaries the indexing data and according to PageRank values Reduce sorts the indexing data and then outputs it to web server. Finally, combining with user interests, web server filters the information that users are not interested and presents the results, which is in the form of webpages, to users.

IV. CONCLUSIONS

This paper analyzes the developing status of search engine and puts forward a new kind of distributed personalized search engine and discusses its key technologies and also gives its theoretical model. Finally under Hadoop platform, we programs and experiments.

Experiment results show that this method can improve the situation faced by search engines, that is slow speed and low precision. The key problem of implementing distributed personalized search engine is to use distributed computing techniques during the course of crawling, indexing and retrieving and obtain user interests. This paper discusses the related technology and algorithmic processes involved to achieve fast and accurate search engine. The next research focus is to use data mining and artificial intelligence technology to build accurate user interest library, improve JobTracker task dynamic segmentation algorithm and distributed webpage gathering strategies in WAN environment. so a more reasonable distributed personalized search engine will be constructed.

ACKNOWLEDGMENT

This work is supported by Gansu Provincial Natural Science Foundation (No.2007GS04864)

REFERENCES

[1] G.E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008,331–338.

[2] WANG Yi-Jie, SUN Wei-Dong, ZHOU Song, PEI Xiao-Qiang, LI Xiao-Yong.Key technologies of distributed storage for cloud computing[J].Journal of Software, 2012,23(4):962−986.

[3] XU Xiao, ZHANG Wei-Zhe, ZHANG Hong-Li, FANG Bin-Xing. WAN-Based distributed Web crawling[J]. Journal of Software, 2010, 21(5): 1067-1082.

[4] Fu ZhongQian, Wang XinYue, Zhou PeiLing etc. Realization of intelligent body on network personalized information filter. Computer Application, 2000, 20(3): 26-29.

[5] Hong Zhang, Yanhong Ma, Qiuyu Zhang. Research on intelligent personalized search engine. ICICT2006:168~172.

[6] WU Wen-Zhong,YI Ping.Application of Distributed Search Engine Based on MapReduce[J]. Computer Systems & Applications. 2012, 21(2):249-250+251.

[7] HU Yu,FENG Jun.Distributed Search Engine Using Hadoop[J]. Computer Systems & Applications. 2010. 19(7): 224-228.

[8] H. B. Liu and V. Kešelj, "Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests," Data & Knowledge Engineering, 2007, 61(2):304-330.

[9] YANG Jing-jing; JU Shi-guang; WANG Xiu-hong. Research of individuation search engine based on web. Computer Engineering and Design, 2008, 29(20):5206~5208.

[10] Zhu Ming. Data Mining. Hefei: China Science & Technology University Press, 2002. 230~231.

[11] Nils J.Nilsson writte. Zheng kougen etc. translate. Artificial intelligence. Beijing: Mechanical Industry Press, 2000. 277~281.

[12] WANG Jun-sheng , SHI Yun-mei , ZHANG Yang-sen.Key technologies of distributed search engine based on Hadoop[J]. Journal of Beijing Information Science and Technology University, 2011, 26(4):53-56+61.