A Preliminary Study of Sort Algorithm for Internet Personal Information Search System

Xu Tiansheng [1, a], Chen Wang

[1] Information college, Capital University of Economics and Business, BeiJing,10007, China

[a]xuts@cueb.edu.cn

Keywords: Web crawler Chinese segmentation Vertical search

Abstract . Today, Internet with a variety of functions has become an indispensable platform for People's living in the community. As a general segmentation and extension of search engine, Vertical Search will deal the webpage information with Chinese word segmentation processing and extract data with the field orientation. Finally, it feedbacks the search results to the user in some kind of form. This paper has a deep analysis of current search engine, by reading relevant literature; I develop a new system about personal information search. This system gets web content through network spider, and then saves the content to MySQL database. The users get links those satisfy the feedback conditions. The system mainly includes two parts, the admin control panel and the user search interface.

Introduction

As the so called Internet, its history will need to track back to the experimental network ARPANET set by the U.S army based on ARPA protocol in 1969.The 90s was the golden age of the rapid growth of Internet. By the end of year 2010, there was 255 million websites on the global scale, during the year after that, the number climbed up to 555 million, which has increase by 117.6 percent1. In 2014 December, China has 18797 block /32 of IPv6 addresses, in addition of 332 million IPv4 addresses2. What's more, the development of search systems has reached unprecedented prosperity, with following features: 1.the size of resource retrieval database is increasing constantly. Commercial search engines keep more than 10 million or ever 100 million web pages in general3.2.vertical searching system has come to emerge4.3.The ability of evaluating the correlation of the results5.4.the usage of automatic classification technology. Due to the rapid development of Internet technology, we are plagued by "information Trek", and are finding it increasingly difficult to locate useful information. In recent years, the domestic retrieval technology has been improving, but in many aspects of the operation is not as good as aboard. Besides, changes of the combination of Chinese vocabulary also have limited the development of Chinese search engine.

This paper transfers the pressure of calculating the weight of user web pages to processing a new formula during the procedure of searching system web pages, which will make search results more accurate, and will enhance efficiency. The background program of the system searches the relevant information on the Internet, the word Web page information, stored in the database, further processing; according to the user input information, the search results to Web page link display. The research methods of this paper:1. Literature research. 2. Theoretical and practical demonstration of the combination. 3. System development tools: HTML+CSS, PHP and MySQL database, etc.

Analysis of vertical search system

Fundamental Search engine, that is, building up a keyword index towards each web page out of those billions of web pages, and then, on the basis of these index, building up a database on top of that, and proceed searching progress in the database. When user need to search a certain word, all pages in the index database contains the content of the keywords will be searched out, and then according to the sorting algorithm, in a certain order output6.

System structure and functions Retrieval system mainly consists of information collection scheduling module, page analysis and access module, index of resource and resource retrieval module, database and log management module and so on. Information search scheduling module mainly uses natural language processing technology on a specific range of priority access, Send hyperlinks to the web access module receive feedback results and analyze the collected data to the web database. According to the hyperlink information in the web database, the administrator uses the HTTP protocol to obtain the information of the link. In the index resource database, the key words in the Internet are preserved; the index resources and the information retrieval module are optimized to speed up the results feedback. The user retrieval module by calling the system retrieval service7, results in HTML format for information feedback.

Realization of control program The main control programs of the vertical search system, that is, the information access and analysis module. Through the SOCKET interface, the results of HTML access analysis are transmitted, and the main control program is inserted into the database according to the pipe interface. The database URL is periodically detected, and the URL is called the main process to access the main process if the database is saved. The main process is the process of web crawler, and saves the relevant information to the database. The process is shown as Fig 1.

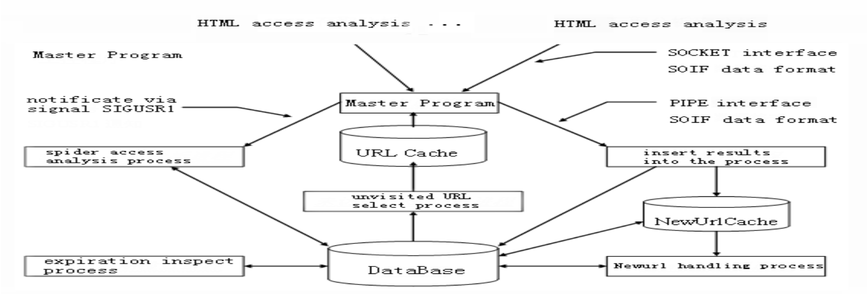
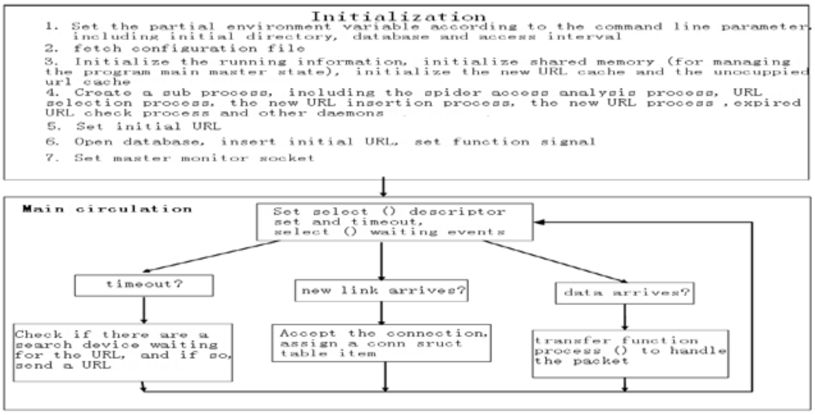


Fig 1 info saving and analysis

The process of main program is as shown in Fig 2.



1. Set the partial environment variable according to the command line parameter, including initial directory, database and access interval
2. fetch configuration file
3. Initialize the running information, initialize shared memory (for managing the program main master state), initialize the new URL cache and the unocuppied url cache
4. Create a sub process, including the spider access analysis process, URL selection process, the new URL insertion process, the new URL process ,expired URL check process and other daemons
5. Set initial URL
6. Open database, insert initial URL, set function signal
7. Set master monitor socket

Main circulation
Set select () descriptor set and timeout, select () waiting events
timeout?
new link arrives?
data arrives?

Check if there are a search device waiting for the URL, and if so, send a URL

Accept the connection, assign a conn sruct table item

transfer function process () to handle the packet

Fig 2 the process of main program

The design strategy of system

Initialize URL The Internet can be considered as a connected graph where the web pages are just like nodes and where the links between web pages are regarded as the edges. Crawler program begins with the initialized URL seed nodes and gets the resources of Internet constantly. It maybe gets less information about website evaluation when it starts to crawl and that leads the system crawling direction to the wrong theme8. System administrators do the job of initializing URL seeds in this system and get webpages by visiting those seeds.

Crawling web Web page crawling is the process of crawlers climb around the edge of the Internet figure. When it accesses to a node, the relevant information of the node will be saved to web database. The crawlers will extract the initialed URLs and put these in the URL working queue, and then traverse the queue. The system will add new URLs those exist in the downloaded webpages to the URL working queue and provide a history table in order to save the flag of whether a certain page is traversed.

The architecture of crawler According to the initialized URL seeds, crawler is able to get links of deeper webpage and save some effective information of it to database. The overall structure is shown in Fig 3

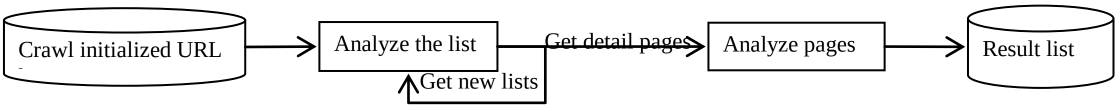


Fig 3 the architecture of crawler

Obtain webpages The Internet resources those including all kinds of file exist in web server are located by URL. As shown in Fig 4, the process might meet the webpages' updating problem. The system should record the date and time when crawlers download the pages. In addition, it manages the content of webpage with MD5 algorithm for checking whether this page has updated.

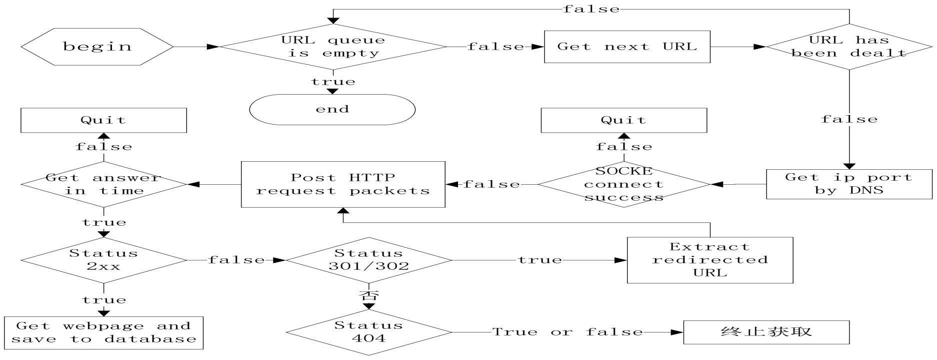


Fig 4 obtain webpages process

Realization of the system

The collation of retrieval results Factors that are considered to sort the results in this system are as following: frequency of page keywords(words_in_page), title keyword(word_in_title), web name(word_in_domain), URL path(word_in_path), meta tag keywords(meta_keyword) and web directory depth(path_depth) and so on. The calculation of page weight is as shown in Eq (1.1).

$$\begin{aligned}\text{weight} = (&\text{words\_in\_page} + \text{word\_in\_title*title\_weight}\\ &+\text{word\_in\_domain*domain\_weight} + \text{word\_in\_path * path\_weight}\\ &+ \text{meta\_keyword * meta\_weight}) \text{ *10 / (0.8 +0.2*path\_depth))}\end{aligned} \tag{1.1}$$

As shown in Equation(1.1), system administrators should set title_weight, domain_weight, path_weight and meta_weight. The path_depth is judged by crawl program.

Inspired by the principle of 80/20(The Pareto Law) 9 in economics, the depth weight of the path is set to 0.2 in this system. The Pareto Law has important practical significance. To avoid spending much time and effort on chores, we have to learn to seize the main contradiction. This paper argues that the darker is the path; their page weights will be relatively lower. But the impact of this factor (path depth) on the page weight cannot be too powerful, so we chose a coefficient of 0.2. At the same time, we regard it as the denominator, not only achieve the goal that the darker is the path, the lower are their page weights, but also realize the aim that the impact of path depth on the page weight is not too serious. It not only seizes the main factors of page weight, but also considers secondary factors.

When system is crawling on the web, firstly, it has to retrieve the keywords in the web pages to determine whether it exists in the above-mentioned sorting factors. If it exists, we should add the weight which the factor corresponds to into the page weight of the keyword. If it doesn't exist , then the weight should be zero. For example: a particular keyword XX exists in the title of a web page A, then set word_in_title of equation 1 to 1; XX does not exist in above-mentioned other factors of the page A, then set the coefficients of the other factors to 0 (that is word_in_domain, word_in_path, meta_keyword all are 0); the depth of the page A is Y, then we can calculate the page weight of page A for the keyword XX.

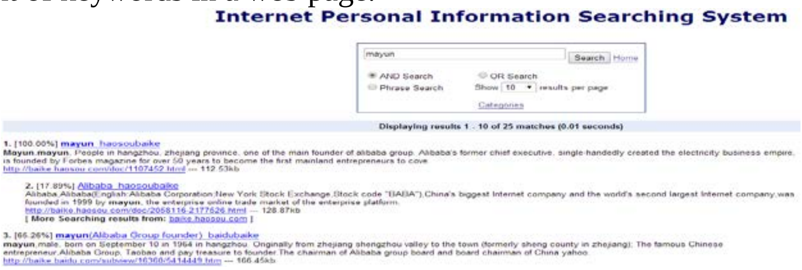
When a user enters a keyword to search, the system firstly retrieves the pages which is associated with the keyword, then, sort the pages which meet the requirements according to the descending of weight. The page whose weight is maximum comes firs. For subsequent pages, calculate the ratio of their weights (maxweight) and that of the first page (that is making the ranking of pages normalized, which was converted into a numerical range of 0-100%). It is shown in the following Eq (1.2).

$weight = number_format($result[$i]['weight']/$maxweight*100, 2) (1.2)

Maxweight indicates the largest weight in the pages which compliance with the requirements, and result array contains the result set meeting the requirements and sorting according to the descending of weight. The function of number_format () in PHP's core- function of String is used to format the number.

Contrast of this system with common search system This system was first proposed vertical search for the Internet personal information that someone searching for a specific type of site information left to achieve in the field for the search function. But ordinary search system to search for information is the smallest unit of the page. Abandon the system description and links to web-based search, using the full text of web page data, word processing and structured approach to the structure of the key units stored in the database, to enhance the degree of association keyword pages, and therefore the user's search results will be closer to their needs.

With the gradual development of web standards toward a direction of HTML5, although some keywords do not display on the interface of a web page, such as words in the Meta tags, but it shares a large component at the time of retrieving information of web pages. New collation proposed in the paper reconsiders common factors impacting web weights, such as header tags, Meta tags and so on. New rules proposed for the factors of path depth achieve the aim that they not only have impacts on page weights, but also the impacts are not too strong. Administrators set up related parameters of the weight, so that when the system is crawling on web pages, it extract the keywords of the web page firstly to determine whether it exists in the factors described in this article, namely weight stored in the database is which have been calculated, which making users can get the feedback at the fastest speed when they search. When the user enters the search field to search, the system retrieves result sets meeting the requirements quickly, and output result sets according to the descending of weight. To calculate the ratio of weights of result sets and the maxweight of them, thus we can obtain correlation of maximum weight of keywords in a web page.


Internet Personal Information Searching System
mayun
Search
Home
● AND Search
○ OR Search
○ Phrase Search
Show 10 ▾ results per page
Categories

Displaying results 1 - 10 of 25 matches (0.01 seconds)

1. [100.00%] mayun haosoubaike
Mayun mayun. People in hangzhou, zhejiang province. one of the main founder of alibaba group. Alibaba's former chief executive, single-handedly created the electricity business empire, is founded by Forbes magazine for over 50 years to become the first mainland entrepreneurs to cove
http://baike.haosou.com/doc/1107452.html --- 112.53kb

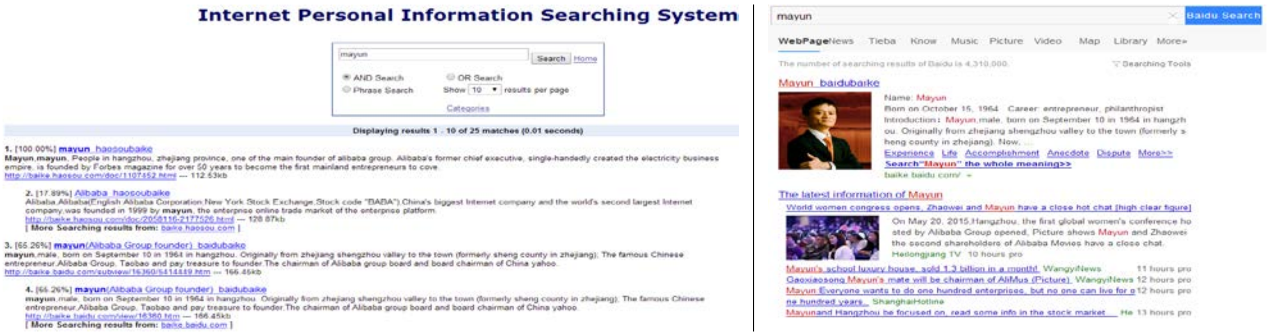
2. [17.89%] Alibaba haosoubaike
Alibaba Alibaba(English:Alibaba Corporation New York Stock Exchange Stock code "BABA") China's biggest Internet company and the world's second largest Internet company was founded in 1999 by mayun. the enterprise online trade market of the enterprise platform.
http://baike.haosou.com/doc/2058116-2177526.html --- 128.87kb
[More Searching results from: baike.haosou.com]

3. [66.26%] mayun(Alibaba Group founder) baidubaike
mayun,male, born on September 10 in 1964 in hangzhou. Originally from zhejiang shengzhou valley to the town (formerly sheng county in zhejiang); The famous Chinese entrepreneur,Alibaba Group, Taobao and pay treasure to founder The chairman of Alibaba group board and board chairman of China yahoo
http://baike.baidu.com/subview/16360/5414449.htm --- 166.45kb

Fig 5 results of searching “mayun” of the system

As is shown above in Fig 5, we can get the weight of “mayun” in haosoubaike is bigger than that in baidubaike by the Eq (1.1). Through click more results, the user can get a message that the higher quality Web page is just some previous links both in haosoubaike and baidubaike. So the system can get some relative info about somebody quickly for the user.

In addition, the system has sorted the website, so users can get information which they have interstate in about somebody through the detail web. As shown in Fig 6, we can see the fact that this function that searching in the fields they are interested in has been achieve, but baidu cannot do this. What’s more, users can also search info by advanced searching options such as “and searching”, “or searching” and “phrase searching”. By “and searching”, the results should include all the searching keywords; “or searching”, the results maybe just about one of the searching keywords; and through “phrase searching”, the results is just about the whole searching keywords.



Internet Personal Information Searching System

Displaying results 1 . 10 of 25 matches (0.01 seconds)

1. [100.00%] **mayun** haosoubaike
Mayun,**mayun**. People in hangzhou, zhejiang province, one of the main founder of alibaba group. Alibaba's former chief executive, single-handedly created the electricity business empire, is founded by Forbes magazine for over 50 years to become the first mainland entrepreneurs to cove
http://baike.haosou.com/doc/1107452.html --- 112.53kb

2. [17.89%] Alibaba_haosoubaike
Alibaba,Alibaba(English Alibaba Corporation,New York Stock Exchange,Stock code "BABA").China's biggest Internet company and the world's second largest Internet company,was founded in 1999 by **mayun**, the enterprise online trade market of the enterprise platform.
http://baike.haosou.com/doc/2058116-2177526.html --- 128.87kb
[**More Searching results from:** baike.haosou.com]

3. [65.26%] **mayun**(Alibaba Group founder)_baidubaike
mayun,male, born on September 10 in 1964 in hangzhou. Originally from zhejiang shengzhou valley to the town (formerly sheng county in zhejiang). The famous Chinese entrepreneur.Alibaba Group. Taobao and pay treasure to founder.The chairman of Alibaba group board and board chairman of China yahoo.
http://baike.baidu.com/subview/16360/5414449.htm --- 166.45kb

4. [65.26%] **mayun**(Alibaba Group founder)_baidubaike
mayun,male, born on September 10 in 1964 in hangzhou. Originally from zhejiang shengzhou valley to the town (formerly sheng county in zhejiang). The famous Chinese entrepreneur.Alibaba Group. Taobao and pay treasure to founder.The chairman of Alibaba group board and board chairman of China yahoo.
http://baike.baidu.com/view/16360.htm --- 166.45kb
[**More Searching results from:** baike.baidu.com]

The number of searching results of Baidu is 4,310,000

Mayun_baidubaike
Name: Mayun
Born on October 15, 1964 Career: entrepreneur, philanthropist
Introduction: Mayun,male, born on September 10 in 1964 in hangzhou. Originally from zhejiang shengzhou valley to the town (formerly sheng county in zhejiang). Now, ...
Experience Life Accomplishment Anecdote Dispute More>>
Search"Mayun" the whole meaning>>
baike.baidu.com/

The latest information of Mayun
World women congress opens, Zhaowei and Mayun have a close hot chat [high clear figure]
On May 20, 2015,Hangzhou, the first global women's conference hosted by Alibaba Group opened, Picture shows Mayun and Zhaowei the second shareholders of Alibaba Movies have a close chat.
Heilongjiang TV 10 hours pro
Mayun's school luxury house, sold 1.3 billion in a month! WangyiNews 11 hours pro
Gaoxiaosong Mayun's mate will be chairman of AliMus (Picture) WangyiNews 12 hours pro
Mayun Everyone wants to do one hundred enterprises, but no one can live for one hundred years ShanghaiHotline 12 hours pro
Mayunand Hangzhou be focused on, read some info in the stock market He 13 hours pro

Fig 6 the difference result of searching “mayun” between this system and baidu

Users can get more results about the searching keywords by click “more results”. The results can be got by clicking “more results” from some certain website. And the weight is recalculated that is to say the weight is just based on this website.

Forecast

With the development of network hardware and software, we need to adjust the crawl search strategy to realize the intelligent transformation between Chinese spell and characters for making the judgment of website more accurate and design some more efficient program to deal with Chinese word segment in order to achieve a better retrieval effect. What we need to do when deploying this system to Internet is that we provide the user a system interface. Display the search results to the user, on the one hand, further optimize web page sorting algorithms, and on the other hand module layer for personal information. These need to perfect the system in the further future.

Acknowledgements

This work was supported by Beijing philosophy and social science project (15ZHB011 13SHB015) and Level of Scientific Research Improving Project of Beijing Municipal Education Commission 2015.

References

[1] China's network. 2012-01-08. http://www.c114.net/news/52/a666482.html

[2] Lixi An. Web information retrieval technology research based on the topic. Shandong University. 2006. Master's thesis

[3] Nie Song. the subject research of search engine with functions of automatic classification. Tianjin University.2004. Master's thesis

[4] Li Liang, Meng Yingjie. Discussion of topic search engine.2003-03-14: The search engine and web information mining academic seminar

[5] Huang Shenggen. The research and design of intelligent vertical search engine.Chongqing University.2010,4

[6] The 80/20 Rule . http://baike.haosou.com/doc/4416385-4623713.html