The principle of a fulltext searching instrument and its application research

Wen Ju Gao [1, a] , Yue Ou Ren [2, b] and Qiu Yan Li [3,c]

[1]Department of Electronic Engineering, Changchun Institute of Engineering Technology, China

[2]Department of Electronic Engineering, Changchun Institute of Engineering Technology, China

[3]Department of Electronic Engineering, Changchun Institute of Engineering Technology, China

[a]gaowj913@sina.com, [b]renyueou@sina.com, [c]liqiuyan@sina.com

Keywords: Principle ; Fulltext; Apache Lucene ; Video ; Index.

Abstract. With the extension of the internet and higher requirement of the retrieval in terms of speed and accuracy, the requirement of the hardware of the search engine is also increasingly rising. On the other hand, network video is now very popular. Network video is excellent in both picture and word, can be played online, and also have the quality of both simpleness and convenience, which are all the incomparable advantages over the words on Webpages. Therefore, network video websites mushroom in large quantity. At present, there are dozens of relatively large domestic network video websites and a lot more are flourishing on net. However, in order to find out the needed video, it would be very inefficiency if visiting those websites one by one, and very limited or no video would be found if visiting just a single website. Based on the principle and procedure of the fulltext search engine Apache Lucene, a new index method is designed for the video information search. It solves the problems such as unclear search engine order or t he limitation in the local area networks.

Preface

According to statistics, in 2003, more than 5000 websites can be visited openly and the total capability of the websites is about 167TB. [1] At the same time, the increasing speed of the information on internet is more stunning. According to one survey, the total number of the websites increases by 108.6% compared with the corresponding period in 2003. [2]In 2007, netizens in the whole world are up to 1.2 billion—more than 1/5 of the population in the world. [3].

At present, there are dozens of relatively large domestic network video websites and a lot more are flourishing on net. However, in order to find out the needed video, it would be very inefficiency if visiting those websites one by one, and very limited or no video would be found if visiting just a single website. Even the famous search engines such as Baidu, can not list the videos for the user in the order of importance and relevance.

Video Information Retrieval System

Generally speaking, system of network video search engine is divided into two parts—video capture system and video information retrieval system. Video capture system is a relatively mature now, which will not be stated here. Video information retrieval system will be interpreted in this paper.

The Summery of Lucene. Indexer and retrieval device are the two important parts in network video information search engine. In order to realize the two functions, Lucene structure is adopted. Some improvement and optimization are made, which make ….meet the demand of the system. Lucene is one of the sub programs among the project group of the Software Foundation. It is an open source full text retrieval engine kit. It's not a complete full text retrieval engine, but a framework of the full text retrieval engine. Fulltext retrieval is a technology which takes text data as its main disposal object, based on full text indexing, using natural language to retrieval. As an open source program, Lucene has been

used extensively, because of its trait of open source, superior indexing structure, and outstanding system framework. It has aroused great repercussions among open source community since it was presented to the public. Developers not only use it to construct concrete full text retrieval application, but also integrate it into many kinds of systematic softwares. Even some commercial softwares use it as the core of the subsystem of the inner full text retrieval. To take the well-known website of the Software Foundation as an example, it uses Lucene as the full text retrieval subsystem. Lucene is also used as the "help" subsystem in the full text indexing engine in the company's open source software. [4].

As to video information retrieval, system will write the attributive information of the videos in the database into the index documents of Lucene regularly. Video information retrieval system is a system used by users directly. In order to find out the required videos, users' first need to input the key words according to their need which will be filtered by the system, then, system will choose and dispose the legal key words. After the disposed key words are retrieved by the Lucene index documents, the required videos will be listed to the user in the order of the importance and relevance of the video. Finally, users can click and watch them. [5]

The Index of Video Information. Based on the video information saved in database, Lucene is used to index these data.

index program will first get the attributive information such as title, synopsis, label, time length, date, browsing times from database; then, the title, synopsis and the label of the video will be disposed through Chinese segmentation process; finally, the result of the Chinese segmentation will be disposed through index analyzer. Because large amount of operations will be involved in the process of generating index, and disk is read and written frequently, in order to accelerate the speed of indexing, first, index segments will be produced in the memory; then, after each segment is completed, all the segments will be combined into the disk. [6]

The Retrieval of Video Information. After getting the index documents of the video information, users need to find out the required video by some means. So that, B-S framework is adopted in the searching module in order that users can search through web browser. [7]

The retrieval device is the main function of the network video search engine and it communicates with users directly, therefore, the design of the retrieval device is a very important part of the search engine. The retrieval device should be designed for the users' maximum convenience and can provide more personalized service.

The specific steps:

(1) Users input the key word they need to inquire, click the search button and send the key word to Servlet server.

(2) To inquire the key word input, and judge the legality of the key word. If the key word is an illegal word concerning content such as reactionary or eroticism, users will be asked to input again. Step 1 is repeated.

If the key word is legal, system will retrieve the index documents according to it. If the retrieval result is zero, step 1 will be repeated; otherwise, go on to step 4.

According to Lucene's retrieval result, database is inquired, and all the video information will be returned.

According to the result returned from the database, the page of search result will be returned to the browser for users. [8]

In order to retrieve the key words more efficiently, arrange the most needed result on the as top of the list as possible and provide more personalized search, the process of the retrieval is optimized in this system. This search engine mainly confronts domestic users, therefore, the key words are mostly Chinese, and the Chinese key words need going through the process of Chinese segmentation. In

addition, users maybe have their own demands on the arrangement of search results, which is to be optimized. [9]

The Design and the Realization of the Index System of Video Information

At present, the retrieval results of search engines are of large quantity, but there is a problem of overloaded information, so that users' requirement can't be reflected clearly. If the search engine can arrange the retrieval results according to the value of the videos and the videos' degree of correlation with the user's key words, users' burden will be alleviated and the efficiency of the retrieval will be improved surely. Therefore, how to arrange the weight of the videos is definitely an emphasis for video search engines. In addition, Chinese is different from western language. Chinese words are not separated by space, but many of them should be separated according to the context. Therefore, how to separate the Chinese key words to make it closer to users' true meaning is an important index to measure a search engine. [10] Lucene index technology based on Java is applied in this video search engine to index the video information in the database and generate the index documents. In order to increase the accuracy of the retrieval results, the Chinese words in the title, synopsis, label and so on will be separated in advance of the index to make them conform to Chinese grammar. Moreover, in order to arrange the retrieval results better, some attributes of the video are processed according to their weight such as the catch time, speed of play, total visit number, the present day visit number of the video. Some videos with relatively late date, bigger visit number, played more smoothly are weighted more, so that these videos will be arranged on the top of the list as much as possible, and at the same time, they are probably the most needed ones to users.

The videos online is mounting rapidly, which result in the considerable capacity of the database. In order to provide the freshest videos for users most rapidly, the system will index the database regularly. At the same time, in order to accelerate the speed of indexing, the multithreading technology of Java is applied. Indexes will be constructed in memory first, then be integrated into the hard disk, which reduces the times of hard disk being read and written, and improves the speed of constructing indexes.

The Effect and the Analysis of the Operation of the System

When users open the searching page, Fig. 3 shows up as follow:

Figure. 3 search box without any key word

Under the default state, the system arranges the search result according to the degree of correlation. As long as the key word shows up in the title, label or synopsis, the search result is up to the mustard. In addition, users can also choose the place where the key word shows up such as title, label or synopsis. At the left of the button of video search, there is a button of advanced search. Users can click it and choose the time range of the video, so that the searching range shrinks greatly, and the degree of accuracy is improved.

In order to make the input of key word more conveniently, when users input part of the key word, the system will give the prompt. For example, if users want to search videos about "Olympics", after"Olympics"is input, the system will prompt automatically more key words relevant to "Olympics", such as"Olympic song", " Olympic knowledge", "Olympic torch" and so on. Users can choose the key word they need with mouse or down arrow key and enter after choosing. Then, the system will search the key word automatically, and show the number of the results on this key word, as picture 4 is showed here:

Figure. 4 the prompt box after inputting ”Olympics”

After we choose”Olympic song) ”, ”Olympic song” will appear on the search box automatically. Search result is showed as Fig. 5:

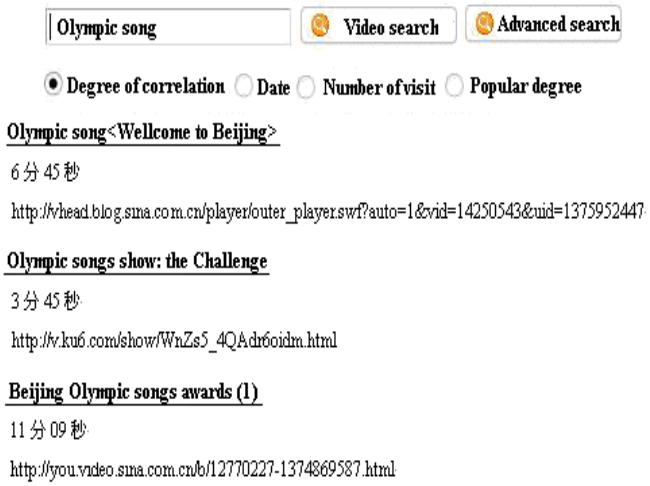

Figure. 5 part of the search result of searching ” Olympic song”

Users can click the item they need on the list and the current page will transfer to the relevant page and the video will be played automatically.

From the search results above, obvious advantages over the other video search engines can be found. Compared with it, other video search engines such as Baidu arrange the search results disorderly. Part of the searching result in Baidu is showed in Fig. 6:

Figure. 6 part of the screenshot of searching result of ”Olympics ” using Baidu

The search engines inside video websites search only within the range of their own websites, which is showed in Fig. 7:

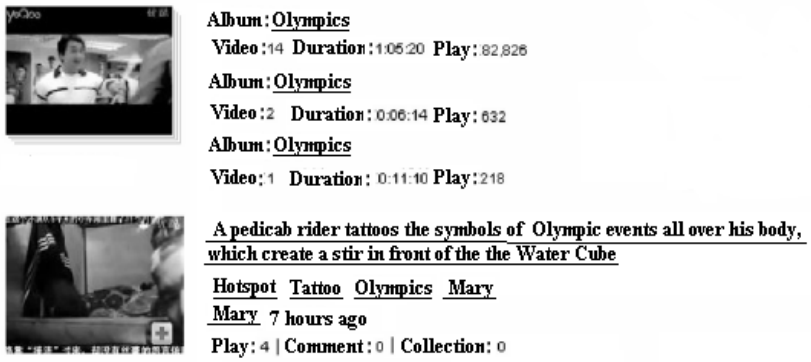

Figure. 7 part of the screenshot of searching result of ”Olympics ” using the search engine in Youku website.

In order to provide better search service, multiple arrangements of the search results are applied, with degree of correlation as the default sort. Users can choose other arrangement according to their own need, such as the date of the video, number of visit, popular degree and so an, which will meet the users' different needs.

Conclusion

With the mounting of the videos online, people are facing the difficulty of searching and visiting their interested video efficiently. In this paper, the method that Apache Lucene is applied to construct a convenient and efficient arrangement of search results is introduced. Obviously, there are much faulty room that can be improved in the future.

References

[1] Information on http://www.sims.berkeley.edu,2009

[2] Information on http://www.cnnic.net.cn,2010.

[3] Information on http://it.21cn.com/itnews/hygc,2013.

[4] Gang Li, Wei Song, Zhe Qiu: Construction of Search Engine, Post & Telecom Press, Beijing (2011)

[5] Gang Li, Wei Song, Zhe Qiu: Conquest or Ajax+Lucene---Construction of Search Engine, Post & Telecom Press, Beijing (2011)

[6] Baowen Xu, Weifeng Zhang: Search Engine and the Technology of Acquiring Information, Tsinghua University Press, Beijing (2013)

[7] Wei Wang, Tieli Zhao, Baixiang Gong, etc: Journal of Changchun University of Technology (Natural Science Edition), 2011.22(02): p. 36-38.

[8] Hong Tan, Junhong Li, Peng Zhou, etc: LUCENEIN ACTION, Electronics Industry Press, Beijing (2009)

[9] Xiaoming Li: Search Engine: Principle, Technology and System, Science Press, Beijing (2013)

[10] Xia Aiyue, The Encryption Method of Network Data and Application Strategy [J] Journal of the Chinese People's Armed Police Force Academy,2012.(8):93-94.