Implementation of Search Engine Based on Java

Yuxiang Hou

Abstract

Search engine is a kind of information retrieval tool based on keyword query, nearly ten years with the popularity of the Internet and the rapid development of the user on the network to find information to provide a very convenient tool. It can be seen that the development and utilization of online information resources for online information search has played a great role in accelerating the exchange of information and promoting the value-added of information. This paper integrates the information mining techniques to provide the novel paradigm for the engine design which will then enhance the overall performance.

Keywords: Search Engine, Java, Information Retrieval, Data Mining

Introduction

Search engine is a kind of information retrieval tool based on keyword query, nearly ten years with the popularity of the Internet and the rapid development of the user on the network to find information to provide a very convenient tool. At present, more types of search engines, more and more brands, richer service, its technology, algorithms, scope of the application, etc. are also different. Traditional search engines indexing through gathering a large number of web sites to provide overall control and network resources retrieval mechanism, so as to help users to easily accurate retrieval to the required information on the Internet. However, in the last decade of search engine development, the traditional search engine has the following defects as the information resources have exploded.

- The application method is incompatible. The different search engine its collection content, the collection method, the emphasis point are different. Regarding the identical user, when its existence different information need that can consider uses many unique search engines.

- Network resource coverage is low. Although some comprehensive search engines at home and abroad are known as millions of Web page collection, Altavista more known as a billion Web page index libraries while with the Internet in the vast number of information resources.

1Nanchang Institute of Science and Technology, Nanchang City, Jiangxi Province, 330108
In general, the frequency and location of words appearing in web pages are represented by the weighted methods. Of course, the actual operation of different search engines is different during the actual operation. Generally, we can analyze this issue from the following aspects. (1) The word frequency weighting methods have absolute word frequency weighting that relative word frequency weighting, counter-word frequency weighting, based on word resolution value weighting and so on. Regarding the sole word search engine, only uses to calculate a word to be possible purely to assign the weight in the homepage appearance frequency but regarding has carries on the logical group to match the function the search engine and then must use other weighing method. For equipping the retrieval model retrieval, retrieval results associated with each term in the retrieval type, and each term in all of the pages in total frequency is different, if the total weight to sort that will cause the result is not relevant. (2) By assigning different weights to the different positions and layouts of the words in the web pages, the relevance of the search results and the search terms is determined based on the weights. (3) On the one hand, the search engine to the best possible relevance to the search results in the front row; the other hand, in the above mentioned 91% of the users through the search engine to find the relevant site, we can see a website to be It is vital that the user click to sort in the search results. In particular, some of the competing sites, in the page search ranking before the search results information that can be regarded as a general competitor.

The Proposed Framework

**Information Mining.** The IDGS system is designed and well developed for automatic collection of Chinese and English technical information on the WWW. IDGS can find information that meets the needs of users and submit them to the users based on the user's target. BP neural network matching technology and statistic-based feature extraction technology, which consists of four parts: pattern extraction module, document acquisition module, pattern matching module and the document library module. In web pages and articles, a lot of data exists in a tabular form. The demand for retrieving key information from web pages and articles is getting higher and higher. In the search for information, it is possible to provide structured
information in the table, and to summarize the highly contrasting information. Compared with
the article, Information description is brief, and the understanding and extraction of the table is
mainly by artificial methods to achieve. Screening and extracting tables for use in
computer-based automated processing is a huge challenge.

Generally speaking, the mining can be separated into the following components. ① pattern
extraction module: According to the user to provide the mining target sample and word
segmentation dictionary, extract the characteristics of the mining targets, and according to the
characteristics of the word frequency to match the pattern matching network to determine the
matching neural network node weight; ② document acquisition module: the use of
representatives of the mining target features and network resource retrieval system to obtain the
initial information node address, and in the accordance with the initial information node address
to run multiple Robot procedures, acquisition source document; ③ pattern matching module:
the use of mode extraction module identified neural network to the collected source document
for pattern matching, will meet the threshold conditions of the document into the basic
document database; ④ document library module: store the excavated documents, and to
provide users with browsing, retrieval and subscription services.

For the mining of the information, the structure can be summarized as follows.

- The keywords in the table content reflect the importance of the table content. We can get
  important data in complex tables by manually setting rules and filtering methods. So by
  manually setting keywords and rules, and can be a flexible match is an important means
to obtain important table data. And manually set the rules than the machine learning to
obtain the rules has a higher efficiency and accuracy.

- Move unless the data form to contain by important form of information processing: to
  form of the complex multiple rowspan or colspan information source code in the manner
  of cell records all the information in the database.

- If you can determine the keyword can be directly retrieved by keyword, otherwise, in the
  same article in the header of the important keywords to retrieve the header information
  and the non-numeric information in the table as a keyword to retrieve.

The Java Techniques. Java in recent years with the development of the network and
popular, it can quickly become a few years in the computer industry's hot spots, which with its
various good features are inseparable. With the establishment of the Java standard, Java
Technology is more mature in the database processing, Java provides a JDBC API for database
developers to develop database applications to provide a standard application programming
interface. There are emerging RMI technologies it presents the concept of remote object
reference, this is the same as the year of the procedural programming language proposed
object-oriented language as revolutionary. In order to Sun Corporation can in the electronic
products to the correlation procedure application, as therefore researched and developed the
Java programming language, in which J2ME realization developed the procedure on the
equipment the function. At present J2ME has two kind of disposition levels, one kind faces the disposition level which the interconnection equipment disposes, it is one kind of model Java hypothesized technology use, it in uses, contained the virtual machine to have uses the function, therefore might had in the memory equipment in some to use. For the Java based engine, we should firstly review the following components.

- Java uses hysteresis binding technology, the object of the domain and the method of memory distribution are implemented by the interpreter to determine. In the bytecode is quoted by the word. For programmers, the direct access to the interpreter's memory is allowed, which is significant for the security of the Internet.

- Java source program after compiling the produce is not the machine code, but the bytecode, then on the platform is used to implement the interpreter to explain, it can be realized in the bytecode compatible with the platform, without having to make any changes on byte.

**Search Engine Design.** The meta search engine must initialize the search engine list before executing the query to select the search engine that needs to be called. There are two ways to choose: system selection and user selection. The former refers to the decision by the system to choose which independent search engine, which is in the meta search engine on the source search engine based on the efficiency of the automatic evaluation of the implementation of the user choice is the user by browsing the search engine list, out of personal preferences and evaluation of the search engine from the main selected independent search engine, the general user is often difficult. In addition, this mechanism can also be achieved on the retrieval time, the number of the results selected users personalized settings. This sort of method is relatively simple, the equivalent of the members of the search engine search results together and then re-select a method of the sorting. This method only improves the recall rate, and for some important information, may be ranked in the more backward position and not easily retrieved by the user while the accuracy is not high.
Based on the figure 2, we present the features of the search engine as the follows. (1) Regarding some inquiry, each member search engine inquiry result speed is dissimilar, in order to reduce the user the standby period, appears the time order sorting returns according to the search engine search result to give the user successively. It is aims at the direct merge speed quite slow question one kind of simple improvement while the sorting result cannot enhance the user the degree of satisfaction. (2) The advantage of the sorting method is that it is simple and fast. The disadvantage is that the abstracts are too simple to return, and the user interfaces of the individual search engines are different, resulting in the same result being returned by different independent search engines. Summary sorting methods often result in search results that return more information than search results, as rather than search results with high ranking results. (3) It will be the members of the search engine to return the results together to re-order, thus disrupting the original search engine sort information, and this information is also very important basis for sorting. Although some members of the search engine sorting method is unknown, but it is certainly according to the query results and the query string the degree of relevance of the sort, but different search engines focus on the different factors. If we can make full use of the members of the search engine sorting information, based on the further synthesis and the ability to further improve the precision.

**Conclusion**

The ultimate goal of the web search engine is to be able to find relevant information based on user needs. It can be seen that the development and utilization of online information resources for online information search has played a great role in accelerating the exchange of information and promoting the value-added of information. This paper presents the novel design paradigm for the search engine and the model hosts the following advantages.
• High calibration rate. On the one hand, the intervention of artificial technology, in the grasp of information by the artificial discovery, identification, in the organization of information by the artificial indexing and with the help of human resources to make such a search engine database construction quality, so the search effect is good. On the other hand a high degree of user participation, under the guidance of the catalog, and gradually clear user's search needs.

• The scientific and standard nature of the classification. For this type of search engine, the classification criteria are one of the important safeguards for providing the quality retrieval services. It is in accordance with the general rules of the people's understanding that both the subject and the classification attribute of the information resources should be met. Otherwise the class is confused and misdirected, and it is easy to fail the inspection.

Reference


