Research on Construction of Network Literature Vertical Search Engine Based on Big Data

Guang Chen

Zhengzhou Institute of Technology, Zhengzhou450052, China

Abstract

With the development of network information technology, mass data continue to emerge to promote the advent of big data. Concerning searching information in the network, copyright theft, infringement, plagiarism and other problems often occur due to declining search efficiency and development and efficient communication of network literature, resulting in severe development situation of network literature. To solve this problem, this paper studies the construction of network literature vertical search engine based on the big data. The model of network literature vertical search engine is built through comprehensive analysis of technologies of vertical search engine and according to the search flow chart, as well as combining keywords and calculation methods, so that the vertical search engine technology is applied to network literature retrieval, aiming at providing higher quality search service for network readers, and promoting healthy development of network literature.

Keywords: Vertical Search Engine, Network Literature, Pagerank Operation, Model Architecture.

1. BACKGROUND

1.1 Introduction

The development and progress of network technologies effectively transform the people's life style and provide convenience for working and living. Online ordering, shopping, ticketing, booking and other services are realized. At the same time, the way of creation transforms from the original using of pen/paper to using of keyboard/mouse, so that much network works emerges (Qiu and Hu, 2015). Network provides convenience for trading, dissemination and use of multimedia goods, greatly promoting publicity and promotion of network literature, and further highlights the advantages of network literature. However, with the in-depth development of network literature and due to lack of related legal provisions and management regulations, it is difficult to safeguard the interest of network products in open network environment, and easy to copy. Piracy is one of the most common problems. Teams or individuals with bad behaviors may copy other’s works and information freely for their interest without permission and spread for profiteering, bringing adverse impact on the copyright owner and relevant contracted website, undermining the credibility of the site while causing huge economic losses (Chen and Lin, 2013). At present, there is no effective way to solve and protect the interests of online literature. In modern society, many copyright protections are confined to papers, journals and magazines, and the network literature is neglected. To solve this problem, it is urgent to select an effective technology to build a new engine system which can strengthen protection of copyright, optimize retrieval methods and meet the needs of users.

1.2 Purpose

In this paper, the idea for construction of vertical search engine is proposed subject to the problems in current network literature, the technology is analyzed in detail, the flow chart is explained, and the effective construction path is discussed (Liu and Xu, 2012). The search engine will collect the effective information in the network by certain strategies and using programs, processing to provide users with search services, and the search results are sorted before display, also known as the common search engine, which is mainly for in-depth exploration in a specific area or industry. After applying it in network literature, the user can automatically crawl and focus on key information using the web crawler, effectively improving efficiency and quality of network pillars, reducing the hardware requirements, not only to meet the needs of users, but to purify the network environment and promote the development of network literature (Tang, 2012).
2. VERTICAL SEARCH ENGINE TECHNOLOGIES

2.1 Flow chart

The main function of vertical search engine is search in particular field, and used widely, such as travel site search engine to provide users with travel information, shopping site search engine to provide consumers with mass commodity information. It can maximize the mining of specific fields for all information resources (Wei, 2011). The vertical search engine consists of the following four parts:

(1) Topical crawler. The topical crawler can grab the paged related to the topic from the Internet according to the strategies, ignoring other pages, and download the grabbed paged to the server.

(2) Web index. The Web index will extract Web information from the pages downloaded to the local server, and establish indexes by weight, Chinese word segmentation, statistical frequency, calculation of weight and other steps.

(3) Page search and sorting. Search the page by the keywords. The search engine will retrieve the pages through the search program to match the keywords, and sort the pages by the sorting algorithm, displaying the sorted pages by the topic and Web information.

(4) User port. The search engine can provide optimized search interface (Zhong and Luo, 2011).

The major difference between the vertical search engine and other common search engines is the web spider crawling strategies and the independent search engine sorting algorithms. The Web spider and sorting algorithms almost determine the quality of a search engine. Figure 1 shows the workflow of vertical search engine technology.

2.2 Key technologies

The search engine grabs information from the network at a regular basis, not when the user searches the network information using the keywords (Fei and Mo, 2014). The search engine collects data from the Internet periodically using the Web spider, storing them in the local Web database, and then establishes the Index database via Chinese word segmentation, information extraction and indexing etc. When the user searches for the keywords, the search engine searches for the keywords and queries the Index database, and then display the results by the page sorting algorithm.
In the vertical search engine system, there are many key technical elements for different functions and roles, with clear division and active cooperation, so as to ensure the smooth operation of entire system (Tong, 2014). This paper focuses on analysis of the core Web spider, also known as the Web crawler, automatically selecting the network information by a law or scripts in data. Its movement trace is like a crawler looking for food in the net, which is called the Web spider. The network information develops rapidly in the past years. Subject to mass data, the search engine shall consider how to download the data from the Internet and store it to the local computer accurately. From which the Web spider came into being, considering which distributed parallelization technology is selected to improve the search efficiency and information extraction speed, and how to filter old/new pages and advertising links. The Web spider is the key and base to search engine technology, and plays a pivotal role. Figure 2 shows the details and workflow of web crawler.

Figure 2. Workflow of Web Crawler

Figure 2 shows the workflow of web crawler based on the path:

1. The Web spider selects some URL as the Seed URL by a method, e.g., including, IP address and/or other methods, and pushes the Seed URL into URL queues to be crawled.
2. The Web spider grabs the Web page link from Internet in order by the URL queue.
3. Perform DNS resolution of the IP address of the link obtained.
4. Download the Web page to the server with the IP address using the download tooling.
5. Deal with the downloaded local Web pages, store them in the Web database for following processing, and record the data which has been downloaded to avoid repeated crawling of Web pages.
6. Extract new URL links from the downloaded Web pages, and match them with the addresses in the URL queue. If they do not match, push the address of the link into the URL queue and update the URL queue.
7. Handle the URL queue in cycle until the queue is empty.

2.3 Statistical methods of Chinese word segmentation

The accuracy of retrieval results provided by Chinese search engine depends greatly on the Chinese word segmentation technology. The current Chinese word segmentation techniques include the following two categories:

Class 1: word segmentation based on string matching. Match the statement to be segmented with the designed words in the dictionary by a specific strategy. If the corresponding word can be found, the statement may be segmented (Zhou, 2015). As a result of forward scanning and backward scanning, the word segmentation methods are divided into forward matching and reverse matching. At present, the word segmentation is one of
the most popular methods, accompanied by morphology, syntax, grammar and other information, simple and efficient.

Class 2: word segmentation based on statistics. Make statistics of occurrences of adjacent characters in the page by syntax, and determine whether they belong to the same word based on the number of occurrences. If the adjacent character co-occurrence frequency exceeds a specified value in the context, it is determined that they belong to the same word. In this way, some new vocabulary is obtained statistically and intelligently with high credibility (Wu and Han, 2013). For example, there are Chinese characters A and B in a statement. If A and B occur much frequently in adjacent, that is, the probability is high. If the probability is higher than the threshold, A and B belong to the same word. Usually the statistical word segmentation needs a dictionary library to filter some unimportant words. This kind of word segmentation requires statistics on the probability of occurrence, featuring low speed but high credibility. The statistical model for affinity degree between different Chinese characters mainly includes mutual information and e test.

Mutual information: Denotes a statement containing Chinese characters J and K. The mutual information algorithm between J and K is

\[
L(J, K) = \log_2 \frac{T(J,K)}{T(J)T(K)}
\]

(1)

Where \(L(J, K)\) denotes the mutual information between J and K. \(T(J, K)\) denote the combined occurrence probability of J and K. \(T(J)\) denote the occurrence probability of J. \(T(K)\) denote the occurrence probability of K. Mutual information can reflect the affinity degree between different Chinese characters. If the value is higher than the threshold, J and K belong to the same word. Where the results of above \(T(J, K)\), \(T(J)\) and \(T(K)\) can be obtained by the following algorithm.

\[
T(J, K) = \frac{N(JK)}{n}, T(J) = \frac{N(J)}{n}, T(K) = \frac{N(K)}{n}
\]

(2)

where occurrence probability of JK, J and K are denoted by \(N(JK)\), \(N(J)\) and \(N(K)\) respectively. \(n\) denotes the total length of Chinese phrase. The mutual information can be used to evaluate the affinity degree of these two Chinese characters. The higher the mutual information, the closer between these two Chinese characters. These two characters meet the condition to form a word; otherwise, they cannot form a word (Zou and Zhao, 2012).

e test: For a Chinese character group RST in order, e test algorithm of Chinese characters S relative to R and T is

\[
e_{R,S}(S) = \frac{N(R/S) - N(S/R)}{\sqrt{\delta^2(N(R/S)) + \delta^2(N(S/R))}}
\]

(3)

where \(N(T/S)\) denotes the conditional probability of T relative to S. \(N(S/R)\) denotes the conditional probability of S relative to R. \(\delta^2(N(T/S))\) denotes the variance of T relative to S. \(\delta^2(N(S/R))\) denotes the variance of S relative to R. It can be concluded that when e test> 0, the character S has a trend to connect with the subsequent character T, the greater the e value, the stronger the trend (Ji and Wu, 2012). When e test = 0, it does not reflect any trend; when e test <0, the character S has a trend to connect with the predecessor character R, the smaller the e value, the stronger the trend.

3. CONFIGURATION OF NETWORK LITERATURE VERTICAL SEARCH ENGINE MODEL BASED ON BIG DATA

3.1 Construction environment

Computer and Operating System: Windows 7, Processor CPU: Inter(R) Core(TM) E6700 3.20GHz, Memory: 8G, Hard Drive: 500G. Table 1 show the detailed Software list.

<table>
<thead>
<tr>
<th>Software name</th>
<th>Edition</th>
<th>Installation folder</th>
</tr>
</thead>
</table>

Table 1 Software Detailed List

380
### 3.2 Configuration of network literature vertical search engine

Configure Java environment variables:

Install the JDK in C:\Program Files\Java\jdk1.7.0_13, and then set the environment variables.

**JAVA_HOME**: C:\Program Files\Java\jdk1.7.0_13

Add `%JAVA_HOME%\bin;%JAVA_HOME%\jre\bin` in PATH;

**CLASSPATH**: .;%JAVA_HOME%\lib\dt.jar;%JAVA_HOME%\lib\tools.jar;

Then configure Nutch:

Add the following information to the configuration tag in file `conf/nutch-site.xml` under `nutch-1.3` folder. `<property>`

```
<name> http.agent.name </name>

<value>MySearch</value>

<description> My Book Search Engine </description>
```

</property>

Add the host site in `conf/crawl-urlfilter.xml` file, which is used to limit the crawling range. Replace `MY.DOMAIN.NAME` in the file `+^http://([a-z0-9]*\="/ qidian.com/` with the site you want to crawl as follows:

```
+^http://([a-z0-9]*\="/ qidian.com/

+^http://([a-z0-9]*\="/ biquge.la/

+^http://([a-z0-9]*\="/ fqxsw.com/

+^http://([a-z0-9]*\="/ 23wx.com/
```

### 4. CONSTRUCTION OF NETWORK LITERATURE VERTICAL SEARCH ENGINE MODEL BASED ON BIG DATA

#### 4.1 Design

To promote the development of network literature and provide convenience for Internet readers, we should actively use the technology of vertical search engine and combine PageRank algorithm to build a new secure search model of network literature based on big data. Figure 3 show the network literature search model flow chart.
Figure 3. Network Literature Search Model Flow Chart

The PageRank calculation in network literature search model includes the following steps:

Step 1: Push the site URL to be grabbed into the stack.

Step 2: Check if the URL is empty. If the URL is not empty, go to the next step; if the URL is empty, the crawling ends, skip to Step 8.

Step 3: Push the top URL out of the stack.

Step 4: Grab the link in the URL and download the Web page.

Step 5: Determine if the link in the URL exists or the Web page exists. (Zhang and Tian, 2016). If it exists, return to Step 1; otherwise, go to the next step.

Step 6: Push the new URL in the stack.

Step 7: Determine if the total number of pages exceeds the threshold. If not, skip to Step 1; otherwise, go to the next step.

Step 8: Get the total number of pages.

Step 9: Conduct iterative computation using improved PageRank formula.

Step 10: Determine if the PageRank tends to be stable. If not, skip to Step 7; otherwise, end.

4.2 PageRank function formula of network literature vertical search engine

The search engine adopts PageRank algorithm, and then conduct offline sorting by the PR value of PageRank algorithm. As the user searches for a keyword, the search engine will display by the RP value of the keyword in the page collection, recommended to the user, and effective improving the response speed. The formula for PageRank is
where \( PR(n) \) denote the weight of page \( n \). \( U \) denotes all pages in the network. \( x_i \) is the page connected to \( n \). \( PR(x_i) \) is the weight of \( x_i \), \( A(x_i) \) is the total number of pages linked to \( x_i \) (Guo and Li, 2016). \( b \) is the damping factor, usually 0.85, mainly used to avoid that the weight of pages without connection becomes 0.

The PageRank can also be calculated by the following formula:

\[
PR(n) = \frac{(1-b)}{U} + b \sum_{i=1}^{U} \frac{PR(T_i)}{A(T_i)}
\]  

Assuming there are four pages in the model. The PR value of the ID is set to: \( PR = 0.3 \). The formula is

\[
PR(1) = \frac{(1-b)}{U} + b \sum_{i=1}^{U} \frac{PR(x_i)}{A(x_i)} = \frac{1 - 0.85}{4} + 0.85 \times (0) = 0.0375
\]

\[
PR(2) = \frac{(1-b)}{U} + b(PR(1)/3 + PR(3)/2) = \frac{1 - 0.85}{4} + 0.85 \times (0.3/3 + 0.3/2) = 0.25
\]

\[
PR(3) = \frac{(1-b)}{U} + b(PR(1)/3) = \frac{1 - 0.85}{4} + 0.85 \times 0.3/3 = 0.1225
\]

\[
PR(4) = \frac{(1-b)}{U} + b\left( \frac{PR(1)}{3} + \frac{PR(2)}{1} + \frac{PR(3)}{2} \right) = \frac{1 - 0.85}{4} + 0.85 \times (0.3/3 + 0.3/1 + 0.3/2) = 0.505
\]

The first round of RP value is calculated. After update of a round of RP value, the importance of the page from high to low is: 4 \( \Rightarrow \) 2 \( \Rightarrow \) 3 \( \Rightarrow \) 1.

Then continue to iterate the evaluation, until all RP values tend to be stable. The higher the weight of page, the more important the page, and the higher the successful search probability (Yu and Qi, 2015). However, there are disadvantages in the PageRank algorithm. For efficient searching, this paper proposes an improvement to the algorithm to make it the core technology for network literature search engine.

As pages may be frequently searched, set the search frequency to \( U \). When the page is not linked, its PR value is the smallest \( \frac{(1-b)}{U} \). To make the time factor affect the PR value, and the new page ranking front, it is necessary to add the time feedback factor \( Q_u \) by the following formula: \( Q_u = f/U \)

Where \( Q_u \) denotes the time feedback factor of the page. \( U \) is the number of times the page searched by the engine, that is, the so-called cycle value. \( f \) is a constant, and the value depends on the minimum value of PR. The PageRank algorithm with the time feedback factor is

\[
PR(n) = \frac{(1-b)}{U} + b \sum_{i=1}^{U} \frac{PR(T_i)}{A(T_i)} + Q_u
\]  

By the above formula, the improved PageRank algorithm is

\[
PR(n) = \frac{(1-b)}{U} + b \left( \sum_{i=1}^{U} \frac{PR(x_i)\cdot Q_u(x_i)\cdot E(x_i)}{A(x_i)} \right) + Q_u
\]  

5. BRIEF CONCLUSION

In this paper, the core technology of vertical search engine is clarified and described after analysis. The statistical method of Chinese word segmentation is analyzed, and the model of network literature vertical search engine based on big data is constructed. The application of the model is realized via optimizing the formula and algorithm of PageRank, so as to provide efficient and convenient searching for online readers, literary creators and literary websites, create a stable and excellent network environment, meet the needs of different people for
network literature, effectively reduce the copyright theft, and lay a solid foundation for sustainable development of network literature.

REFERENCES

Chen T., Lin J. (2013). Comparative analysis of temporal and spatial evolution of online public opinion based on the attention of search engine - Taking Google trend and Baidu index as an example, Information magazine, 32 (03), 7-10+16.


Tong F. (2014). In network search engine caused by the issue of copyright, Economic research Herald, (15), 298-300.

Wei X.Q. (2011). Thoughts on copyright infringement liability of Internet search engine providers, Legal system and society, (05), 80-81.


Zou S., Zhao S.Y. (2012). Research on Algorithm of word segmentation in vertical search engine, Computer technology and development, 22 (02), 131-133+137.