

## **A The Design and Realization of Open-Source Search Engine Based on Nutch**

Wei Tong<sup>1,2</sup> Xiaoyao Xie<sup>1,2,3</sup>

1. School of Computer Science and Technology  
Guizhou University

2. Key Laboratory of Information and Computing Science of Guizhou Province  
Guiyang, China

3. Guizhou Normal University  
Guiyang, China

[xyx@gznu.edu.cn](mailto:xyx@gznu.edu.cn) (corresponding author: Xiaoyao Xie)

### ***Abstract***

*Nutch is an open-source search engine framework based on Lucene Java. On the basis of the web crawlers and NDFS files system Nutch provides, a proficient search engine system can be developed to search, find out, filter, segment information and then provides users with searching service. One of the biggest advantages of it is open-source, in accordance with which some algorithms of search engine and data framework can be developed. Furthermore, by using Nutch, enterprises can invent an appropriate web search engine according to their own characteristics and needs.*

*This paper, through a research into Open-source search engine Nutch, introduces how a common search engine works. It usually includes steps as follows:*

*(1) Pages collection (fetch). The program of collecting pages, by timely collection or incremental collection, chooses the URLs, through which pages are to be visited and then fetched to the local disk by the crawler.*

*(2) Creating index. The program of creating index converts the pages or other files into the txt-document, divides them into segments, filters some useless information and then, creates and assists indexes which are composed by some smaller indexes based on key words or inverted documents.*

*(3) Searcher. The program of searcher accepts user's query words through segmentation and filtering and then divides them into groups of key words, according to which correspondent pages are matched in treasury index. Then, it puts the matches in order by sorting and returns the results to the users.*

*By using Nutch, a search engine which belongs to Guizhou Normal University's website is designed. On account of the fact that there are so many sites under Guizhou Normal University's website, not only the pages but also some other resources like doc, pdf are needed to be indexed. In this sense, adding the text analyzer module to the design based on Nutch's framework, the whole design is composed by the crawler design module, the text analyzer module, the index module and the search module.*

*In these modules, the most important some work need to do is listed as follows:*

*Page Voice Elimination: After getting the content, the pages include a lot of tags and other ad information. It is necessary to eliminate these spasms and get the effective document. Here the program must complete two missions.*

*The sorting search results: The Nutch's sorting model has some limits: First, Web has mass data. The page includes a lot of insignificant and iterant messages which affect the information that users really want. The model cannot deal with these messages well. Second, the precision of the query is not very good, and it does not show the weightiness of the web*

page. Under this situation, we have done some improvement to Lucene's sorting algorithm which is shown as follows.

The improved algorithm:

$$\text{Score}_d = k1 * \text{OldScore} + k2 * \text{PrScore} + k3 * \text{ReScore} + k4 * \text{homePageScore}$$

$\text{Score}_d$ : Record  $d$ 's score.

$\text{OldScore}$ : The  $d$ 's score is calculated by the Lucene's sorting algorithm.

$\text{PrScore}$ : Record  $d$ 's PageRank score.

$\text{ReScore}$ : Record  $d$ 's score when if the document has been queried for a second time.

$$\text{ReScore} = \text{rescore} + (\text{hitNum} - 1) * \text{increment}。$$

$\text{homePageScore}$ : Record the homepage's score.

$K1, K2, K3, K4$  are Weight coefficient  $PR(A) = (1 - d) + d(PR(1) / C(1) + \dots + PR(n)/C(n))。$

PageRank, second query, the home PageScore has optimized the precision of the searching process.

At last, through the improvement of Nutch's sorting algorithm and experiment, it can be found that Nutch is very suitable for working in home-search. After the Crawl Testing and deploying the project into the tomcat, we can see the results from the figure 1.

Key words: Search Engine; Nutch; Lucene;Java Open Source;



Figure 1