Analysis and Improvement of Chinese Index Technology of Open Source Search Engine Nutch

To cite this article: Min Tan and Hong Lan 2019 J. Phys.: Conf. Ser. 1176 022032

View the article online for updates and enhancements.
Analysis and Improvement of Chinese Index Technology of Open Source Search Engine Nutch

Min Tan, Hong Lan*

School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi, China

*Corresponding author e-mail: lanhong69@163.com

Abstract: It is difficult for the computer to correctly understand Chinese language, because Chinese is a very complicated language. Therefore, there are many problems existing in Chinese search technology. This paper focuses on solving the defects of Chinese word segmentation technology. Proposes an algorithm which is named as Chinese word segmentation system based on two-way maximum matching. Applies the algorithm into Nutch search engine, testing results shows that the proposed algorithm improved the abilities of Nutch system in Chinese processing.

1. Introduction

The 21st century is an era of information and network. The continuous smooth and progressive Internet not only brings huge amounts of network information to people, but also makes people confused when they inquire about useful relevant information they need\cite{1}. Search engine is the network tool to satisfy people's network information search\cite{2,3}. It is the guiding principle of Internet information query.

Nutch is a complete web search engine based on Lucene\cite{2}, developed in Java\cite{3} language, and designed with plug-in mechanism. By extending extension points, users can develop their own plug-ins. Its appearance makes grabbing, indexing and retrieval integrated, and has the characteristics of good scalability and open source.

Chinese word segmentation\cite{4, 5} is a key technology of Chinese search engine. The quality of word segmentation directly affects the accuracy and efficiency of search. Nutch is developed for English, and only for one word Chinese. It can not meet the actual needs in the efficiency and accuracy of Chinese word segmentation. To achieve better search and query of Chinese information, it is necessary to add Chinese word segmentation method in Nutch.

By studying Nutch's architecture and workflow, this paper proposes an improved Chinese word segmentation system based on bidirectional maximum matching algorithm to optimize the effect of word segmentation and improve Nutch's Chinese processing ability. Nutch improves Chinese word segmentation only by extending Chinese word segmentation points. By comparing the improved algorithm with the original Nutch word segmentation\cite{5} and Lucene dichotomy\cite{6}, it can be seen that the improved algorithm has a significant improvement in word segmentation effect.

2. Nutch

Nutch is an open source Java search engine based on Lucene, which is dedicated to providing users with the services of agile development of Web search engines\cite{4, 7, 8}, providing all the tools needed to develop search engines, including full-text search and Web crawlers\cite{8}. 
2.1. Nutch overall structure
Nutch is generally divided into three main parts: crawling, indexing and searching. The relationship
between these parts is shown in Figure 1. Web DB is the set of URLs that Nutch initially runs; Fetcher
is the crawler used to grab web pages, which is commonly called Crawler; indexer is the part used to
build indexes, which will generate index files and store them in the system; searcher is the query,
which is used to complete the search of a term and return the results.

![Figure 1. Nutch overall structure](image)

2.2. Nutch spider
Nutch system contains a powerful web spider, the core of which is the Crawl tool. This tool
automatically downloads pages according to the pre-set list of entry URLs until the system's default
stop conditions are met. Crawl itself is a combination of other tools for downloading web pages. With
these underlying download tools, more flexible web page download tasks can be achieved.

Crawler's working principle: Firstly, Crawler generates a set of URLs called Fetchlist according to
WebDB, then the download thread Fetcher crawls the web page according to Fetchlist. If there are
many download threads, then it generates many Fetchlists, that is, a Fetcher corresponds to a Fetchlist.
Crawler then updates WebDB with the crawled pages, generates a new Fetchlist based on the updated
WebDB, which contains uncovered or newly discovered URLs, and then restarts the next round of
crawling. This cycle process can be called a "generation/capture/update" cycle.

3 Chinses Words Divided Syncopation Technology
The process of reorganizing continuous self-orders into word sequences according to certain norms is
called word segmentation. The definition of word segmentation in Modern Chinese Word
Segmentation Standards for Information Processing is: the process of dividing Chinese into word
segmentation units according to specific norms and the need of information processing. For English
word segmentation, since English is word-based, sentences can be well analyzed simply by using
spaces as delimiter. Unlike English, Chinese word segmentation is to divide the sequence of Chinese
characters into meaningful words. Because there is no space between words in Chinese sentences, and
the combination of the two words seems to be a word in a sentence may not necessarily be a word, in
order to recognize the words in Chinese sentences, the computer must adopt new technology different
from English word segmentation.

Chinese word segmentation technology belongs to the category of natural language processing
technology. It is the first step in the process of semantic understanding. It extracts the core words of
sentences for the use of semantic analysis module. In the process of word segmentation, how to
properly provide enough words for the analysis program to process, and how to complete the process
by computer, this process is called word segmentation algorithm.

It is generally believed that there are three main forms of Chinese word segmentation methods,
including dictionary matching based method, semantic understanding based method and statistical
method[^5][^9]. The word segmentation method based on dictionary matching is called mechanical word
segmentation. Mechanical word segmentation has been well developed in the early stage of Chinese information processing. At present, the mainstream is automatic intelligent word segmentation.

4 Nutch Chinese Index Technology and Improvement

4.1 An Improved Nutch Chinese Word Segmentation Algorithms

Writing a participle for Nutch must be extended from the org. apache. nutch. analysis. Nutch Analyzer extension point. This time, we will use the dictionary-based two-way maximum matching algorithm to improve Nutch segmentation.

Two-way matching segmentation is mainly to compare the positive maximum matching word segmentation and the reverse maximum matching word segmentation, and obtain the best segmentation effect of the two words. The flow chart of the algorithm is shown in Figure 2.

![Figure 2. Flow chart of bidirectional maximum matching algorithm](image)

The flow chart of the forward maximum matching algorithm is shown in Figure 3. The reverse maximum matching algorithm is similar to the flow chart of the forward maximum matching algorithm.

![Figure 3. Flow chart of forward maximum matching algorithm](image)
The specific implementation steps are as follows:

1. A new SlbAnlyr class is created in the package org.apache.nutch.analysis. By implementing the tokenStream (String field Name, Reader reader) method of this class, the extension of Nutch to Chinese participle is realized.

2. A new SlbTokenizer class is created in the package of org.apache.nutch.analysis, which is mainly used to analyze text documents and is called by the SlbAnalyzer class.

3. Modify the new FastCharStream class in the org.apache.nutch.analysis package. Because this class tracks the character position of Token in the input, it must call the Chinese word segmentation method to segment Chinese words, and then record the character position of Token according to the result of the word segmentation.

4. Create a new LoadDic class in the package org.apache.nutch.analysis. Since bi-directional maximum matching algorithm is a mechanical word segmentation method, which needs the support of dictionary, the effect of word segmentation is determined by dictionary, so this kind of algorithm will read dictionary files and then put them back in vector.

5. Create a new MaxMatchSegment class in the package org.apache.nutch.analysis. This class is the realization of bidirectional maximum matching algorithm. This class mainly carries on the forward maximum matching and the reverse maximum matching to the string, and then compares them, returns the result of less segmentation, and if the number of segmentation is equal, takes the result of reverse maximum matching.

6. New ZxSegment class and FxSegment class in the package org.apache.nutch.analysis. The forward maximum matching algorithm and the reverse maximum matching algorithm are implemented respectively. The segmentation results are put into the list, which is called by the MaxMatchSegment class.

7. Modify the NutchDocument Analyzer class in the org.apache.nutch.analysis package to call the SlbAnalyzer class, so as to improve the Chinese word segmentation of Nutch.

8. In src/java/org/apache/nutch/analysis/NutchAnalysis.jj, \(<\text{SIGRAM}: \text{CJK}\rangle\) is modified to \(<\text{SIGRAM}: (\text{CJK})^+\rangle\), and then NutchAnalysis.jj is compiled with JavaCC to overwrite the original Java class.

9. Re-compile nutch-1.0 with ant, and overwrite the jar and job files in the build over the corresponding files in / nutch-1.0. So far, the Chinese word segmentation of Nutch has been improved and deployed.

5 Experiments and Tests
Through the above analysis of Nutch Chinese word segmentation principle and code compilation, we deploy Nutch Chinese word segmentation system. After configuring, we test the improved Nutch based on bidirectional maximum matching algorithm.

5.1 Word Segmentation Effect Test of Improved Algorithms
The results of the forward maximum matching word segmentation test are shown in Figure 4(a), the test results of reverse maximum matching segmentation are shown in Figure 4(b), The results of the two-way matching word segmentation test are shown in Figure 4(c).
5.2 Web crawling effect test

The whole process of crawling web pages includes establishing crawl directory, configuring crawl-urlfilter.txt file, modifying nutch-site.xml configuration file, setting HTTP proxy name, and finally running crawl to crawl web pages [4, 10, 11].

The system is tested on campus network, so crawl command is adopted. Run cygwin software, go to nutch-1.0 root directory, and enter bin/nutch crawl weburls.txt-dir local-depth 3-topN 10-threads 5 to return. The Cygwin crawl page command window is shown in Figure 5.

When the crawl is completed, a local folder is generated in the nutch-1.0 directory, and a log file named logs is also generated. Using this log file, you can analyze any errors you may encounter, and the files generated during the web crawling process are shown in Figure 6.

After web crawling, Nutch builds an index, which can be viewed by Luke. The index built by Nutch before and after improvement is shown in Figure 7 (a-b). In addition, in order to verify the experimental results of the improved algorithm, the Chinese word segmentation package CJK Analyzer of Lucene is used for binary segmentation of Chinese word segmentation. The index established by binary segmentation is shown in Figure 7 (c).

(a) before improvement  (b) after improvement        (c) dichotomy

Figure 7. Index established

Compare results, the improved algorithm has a better segmentation effect than before. Nutch has a poor ability to segment Chinese phrases before the improvement. Then Nutch can accurately segment two-word phrases and three-word phrases after the improvement of this algorithm.

5.3 Search testing with Tomcat

Copy the nutch-1.0.war (war is the suffix name of the Java Web project distribution package) in the
nutch-1.0 directory to the Tomcat\webapps directory, modify the configuration file nutch-site.xml, start Tomcat, open the browser, at the address Enter http://localhost:8080/nutch-1.0 in the column to perform the query test[12]. For example, enter the query word “University of Science and Technology” in the search box, and the search result returns as shown in Figure 8.

![Figure 8. Search Results Page](image)

6 Conclusion
Nutch is designed for English, just one-word segmentation in Chinese can be supported. In order to enable Nutch to search and query Chinese information better. Based on the research and analysis of Nutch’s architecture and workflow, this paper proposes an improved Nutch Chinese word segmentation algorithm based on bidirectional maximum matching algorithm through the extension points provided by Nutch. Compared with the original Nutch word segmentation and Lucene dichotomy, the experimental results show that the improved algorithm has a significant improvement on Chinese word segmentation, which verifies that the algorithm can enhance Nutch's ability to process Chinese words.

References
[1] Li Hui. Database Technology and Application [M]. Tsinghua University Press, 2016.
[2] Zhou Dengpeng, Xie Kanglin. Lucene search engine [J]. Computer Engineering, 2007,33(18): 95-6.
[3] Geng Xiangyi, Zhang Yueping. Java 2 Practical Course [M]. Tsinghua University Press, 2012.
[4] Wang Xuesong. Lucene+nutch search engine development [M]. People's Posts and Telecommunications Press,2008.
[5] Gaofei, Liu Yun. Realization of Chinese Word Segmentation Method in Nutch [J].Network Security Technology and Application, 2008, 9): 73-4+90.
[6] Liu Yang. Realization of Chinese word segmentation in dichotomy in Nutch 0.9 [J]. Computer Age, 2009, 4): 171-2.
[7] Huang Bo. Research and Application of Subject Search Engine [D]; Chengdu University of Technology, 2007.
[8] Gao Wenju, Li Xiaowei, Sun Chunyan, et al. Research on the Principle and Process of Apache Lucene Engine Based on Full Text Retrieval [J]. Journal of Changchun University of Technology: Natural Science Edition, 2008, 29 (4): 424-7.
[9] Realization of JE Chinese word segmentation in Cai Xiaoyan, Kou Yingzhan, Shen Wei, et al. Nutch-0.9 [J].Science and Technology and Engineering, 2008, 8(17): 4881-4.
[10] Deng Ziyun, Zhang Zhi. JSP Network Programming: From Foundation to Practice [M]. Electronic Industry Press, 2005.
[11] Guo Zhen. JSP Programming Course [M]. People's Posts and Telecommunications Press, 2008.
[12] JASONBRITTAINT, IANF. DARWIN. Tomcat Authoritative Guide [M]. China Electric Power Publishing House, 2009.