Chinese Tourism Information Search Platform based on Cloud Computing

Zhao Huan[1,a], Chen Xi[2,b]

[1]Education Technology Center, Beijing International studies University, Beijing 100024, China

[2]Education Technology Center, Capital University of Economics and Business, Beijing 100070 China

[a]email: zhaohuan@bisu.edu.cn, [b]email: chenxi@cueb.edu.cn

Keywords: Nutch; Hadoop; Solr; Chinese word segmentation; cloud computing

Abstract. Nutch, Solr and Hadoop, three of them are open source applications, which Nutch is a superb web crawler, Hadoop is a cloud platform and Solr can use crawled data and offer word class searching. Nutch only provides one mechanism which segments Chinese sentences into some single characters so that Chinese word cannot be analyzed and processed. This paper proposes a method of Chinese word segmentation in Solr and builds a high performance distributed search engines by integrating Nutch into Hadoop, and finally use Solr to build tourist information search platform.

Introduction

With the popularization of information technology, we can obtain all the information we need from the Internet. Recently there is a fast growing demand for tourist information. However, because of the massive growth in tourism information and the large scale use of twitter, e-commerce and SNS [1], how to extract effective information from huge data in tourism information has attracted people's widespread attention.

We need to segment tourism information obtained by Internet for better search results [2]. At present, many scholars are doing a lot of research work in the field of Chinese word segmentation and achieve many results. Words, not Characters and sentences are basic terms in Chinese language. Due to specification of Chinese language, we can understand Chinese text well only after we segment Chinese into some words correctly. In this paper, we introduce a tourism information collection and words segmentation procedure which is implemented on Hadoop platform.

Structure of the tourism information search platform

Tourism information search platform is divided into three parts, namely crawling of tourism information, Chinese word segmentation of tourism information and tourism information search server, as shown in Fig.1.

Implementation of crawling for distributed tourism information

Nutch is an open source distributed search engine written in Java, including crawler and searcher. By using it, we can crawl web pages in an automated manner [3]. Hadoop which Apache develops is an open source cloud platform similar to GFS and MapReduce of Google. It is a framework that allows for the distributed processing of large data sets across clusters of computers in a highly-available way.

1 Cluster environment

Nutch can run on a single machine, but gains a lot of its strength from running in a Hadoop cluster when working with a large data source.

1.1 Server environment

We should edit every server hostname, compare every server's IP with its hostname, install openssh-server and configure public key and private key in each node and install JDK and

configure JAVA_HOME;

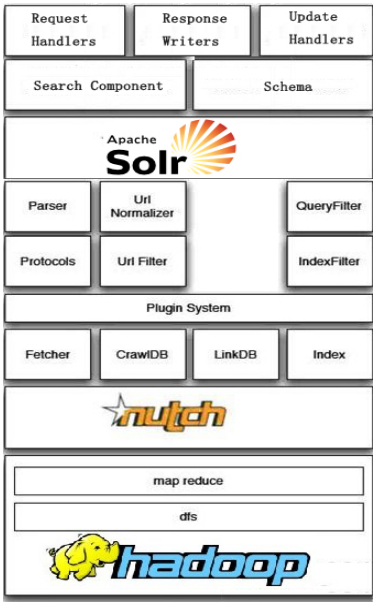

Fig.1. Structure for tourism information search platform.

1.2 Configuration of Hadoop and Nutch

- Configuration of Hadoop

Typically one machine in the cluster is designated as the NameNode and another machine the as JobTracker, exclusively. The rest of the machines in the cluster act as both DataNode and TaskTracker [4]. To configure the Hadoop cluster, the environment should be configured in master as well as slaves. There are four configuration files, namely Hadoop-env.sh core-site.xml mapred-site.xml and hdfs-site.xml to configure so that Hadoop cluster can work.

(1)Hadoop-env.sh

Edit the file Hadoop-env.sh to define some parameters, like Hadoop_HOME, JAVA_HOME and Hadoop_LOG_DIR.

(2)core-site.xml

```
<name>fs.default.name</name>
<value>hdfs://master:9000</value>
……
```

(3)hdfs-site.xml

```
<name>dfs.name.dir</name>
<value>/Nutch/filesystem/name</value>
……
<name>dfs.data.dir</name>
<value>/Nutch/filesystem/data</value>
……
<name>dfs.replication</name>
<value>1</value>
……
```

(4)mapred-site.xml

```
<name>mapred.job.tracker</name>
<value>master:9001</value>
<name>mapred.map.tasks</name>
<value>2</value>
……
<name>mapred.reduce.tasks</name>
<value>2</value>
……
```

<name>mapred.system.dir</name>
<value>/Nutch/filesystem/mapreduce/system</value>
……
<name>mapred.local.dir</name>
<value>/Nutch/filesystem/mapreduce/local</value>

- Configuration of Nutch

(1)The file Nutch-site.xml serves as a place to add custom crawl properties that overwrite Nutch-default.xml. The only required modification for this file is to override the value field of the http.agent.name, and edit Nutch-site.xml in all nodes:

```
<property>
    <name>http.agent.name</name>
    <value>Tourism search</value>
</property>
```

(2)The file regex-urlfilter.txt will provide regular expressions that allow Nutch to filter and narrow the types of web resources to crawl and download, and edit regex-urlfilter.txt in all nodes:

```
+^
#skip everything else
-.
```

2 crawling data in a cluster

After running Hadoop, add 'HADOOP_HOME/bin' into environment variable, and then make Nutch worked in Hadoop. bin/Hadoop fs -put urldir urldir, the first urldir is local directory, which includes some tourism information links, and it can be used as entry point for Nutch crawler, as shown in table 1. The second urldir is storage path of HDFS.

Table 1 entry point for Nutch crawler

http://travel.sina.com.cn	Sina Travel
http://travel.163.com	163 Travel
http://go.qq.com	Tencent Travel
http://travel.sohu.com	Sohu Travel
http://www.yododo.com	Yododo Travel
http://www.uzai.com	Uzai Travel

Now we are ready to initiate a crawl and run the following command in the directory of NUTCH_HONE/runtime/deploy: $bin/crawl conf/urls/seed.txt crawl http://master:8983/Solr/ 90

Implementation of Chinese word segmentation in tourism information

Solr is the popular, blazing fast open source enterprise search platform from the Apache Lucene project [5]. We take all data from Nutch crawling export to Solr, and segment all data with Chinese word segmentation toolkit. IK Analyzer is an open source word segment module implemented by Java. This paper proposes to use IK Analyzer Chinese word segmentation toolkit based on dictionary to transform Nutch, and implement Chinese word segmentation on tourism information. The process is exemplified as follows [6].

1 Implementation of Chinese word segmentation

- Configure Tomcat and copy files from path of Solr to path of Tomcat.
- Install Solr-work path

```
<Context docBase="$TOMCAT_HOME/webapps/Solr" debug="0" crossContext="true" >
<Environment name="Solr/home" type="java.lang.String" value="$TOMCAT_HOME/Solr" override="true" /></Context>
```

- Configure IK Analyzer

```
<fieldType name="text_ik" class="Solr.TextField">
    <analyzer type="index" isMaxWordLength="false" class="org.wltea.analyzer.lucene.IKAnalyzer"/>
    <analyzer type="query" isMaxWordLength="true" class="org.wltea.analyzer.lucene.IK
```

Analyzer"/> </fieldType>

2 Comparison and Analysis of word segmentation

After crawling data, we can use Luke to analysis the Chinese word segmentation, as shown in fig. 2 and fig. 3. Fig. 2 is the results before adding IK analyzer and fig. 3 is the results after adding IK analyzer. As shown in fig.2, which segments Chinese into some single characters and doesn't conform to Chinese language habit of segmenting Chinese into some words [7]. However, fig. 3 shows that the word segmentation result of IK Analyzer module more conforms to Chinese language habit.

Implementation of Tourism Information Search Server

After configuration of Tomcat and Solr, we can run Solr's search interface. First, files from directory of dist and contrib should be copied to directory of Solr in Tomcat, and then edit Solrconfig.xml. Lastly, try it out athttp://master:8983/Solr/browse.

Freq	Field
127143	text
120732	text
119821	text
117080	text
117074	text
116891	text
115612	text
114965	text
114521	text
113109	text
111922	text
109881	text
107464	text
107402	text
107398	text

Fig.2. Before add IK analyzer

Freq	Field
130769	text
129862	text
129481	text
127787	text
127604	text
127270	text
126161	text
125765	text
118489	text
117527	text
122375	text
121941	text
119747	text
119703	text
119604	text

Fig.3. After add IK analyzer

Conclusion

This paper designs an efficient, reliable and scalable Chinese search engine by using Nutch and Hadoop, and then segments the searching data by using Chinese word segmentation, and finally uses Solr to build a search server. To evaluate the effect, we build a cluster with six servers and deployed Nutch and Solr on the cluster. Experiments show that after joined the segmentation module, segmentation of Chinese word is better, moreover users can updated dictionary according to their own needs, which improves the segmentation module functions, and has large significances on the theory and application of tourism information search.

Acknowledgement

In this paper, the research was sponsored by the Science and technology planning item of Beijing Municipal Commission of Education (Project No. SQKM201410031001) .

References

[1]Yang L, Shi Z. An Efficient Data Mining Framework on Hadoop using Java Persistence API[C]// The 10th IEEE International Conference on Computer and Information Technology (CIT-2010). Bradford, UK. USA: IEEE, 2010.

[2]Wang Xue-song. The development of search engine based on Lucene + Nutch[M]. Beijing: People Posts Press,2009.

[3]Zhao Yan-rong, Wang Wei-ping, Meng Dan, et al．Efficient join query processing algorithm CHMJ based on Hadoop[J]．Journal of Software,2012, 23(8):2032-2041．

[4] KHATCHADOURIAN S，CONSENS M，SIM ON J．ChuQL: processing XML with XQuery using Hadoop[C]// Proceedings of the2011 Conference of the Center for Advanced Studies on Collaborative Research．Riverton: IBM Corp．2011: 74-83．

[5]Zhai Feng-wen, He Feng-ling, Zuo Wan-li．A Chinese word segmentation method which combines dictionaries and statistical[J]．Mini micro system,2006,27(9):1766-1771．

[6]Tom Wbite, Zeng Da-dan．The definitive guide of Hadoop(Chinese edition)[M]．Beijing: Tsinghua University Press,2010: 4448．

[7] SRIRAMA S N，JAKOVITS P，VAINIKKO E．Adapting scientific computing problems to clouds using MapReduce[J]．Future Generation Computer Systems， 2012，28(1):184－192．