A Price Comparison System Based on Lucene

Jianxia Chen [1],
School of Computer Science
Hubei University of Technology
Wuhan, P.R.China
chenjianxiawh@gmail.com

Ri Huang [2],
School of Computer Science
Hubei University of Technology
Wuhan,P.R.China
714964653@qq.com

***Abstract*—Lots of online shopping systems (OSS) are proposed and used practically due to the rich opportunities provided by the Internet. The traditional OSS, however, essentially provides basic browsing via category and "advanced" keyword without any analysis. The paper presents a price comparison system of online products to show all the possible prices of products for customers. In particular, the proposed system develops a multithreaded crawler to implement web information crawling, and uses Lucene, a very popular full-text search library, to implement the data indexing and retrieval. The experimental results demonstrate that the proposed system improves shopping efficiencies for the consumers in a flexible and advanced way.**

***Keywords*—search engine; price comparison; crawler, Lucene**

I. INTRODUCTION

Recently, many online shopping systems (OSS) are proposed and used practically due to the rich opportunities provided by the Internet. Also, OSS is being developed to allow consumers to have a more convenient and more interactive platform during their shopping process .In particular, the technology implementation in an evaluation of the same commodities from the different websites can be one of the strongest influencing factors in stimulating consumers shopping motivation.

From the perspective of evaluation about the commodities, the paper designed and implemented an innovative online product price comparison system (PCS), incorporated the popular web search engine technologies, including web crawling, web data extraction, data retrieval. The proposed system provides the consumers the similar products' price information from the different online shopping malls, thus assists the consumers to achieve the traditional shopping goal that "shop around, rational consumption" in a short way. The experimental results demonstrate the proposed system improves shopping efficiencies for the consumers in a flexible and advanced way.

The paper is organized as follows. In Section 2, the paper introduces Lucene technologies in the proposed system. The proposed system PCS infrastructure is introduced in Section 3. In Section 4, the paper discusses the implementation of the proposed system. The paper presents the experimental results in section 5. Finally, we give some conclusions and future work in Section 6.

II. LUCENCE TECHNOLOGY

Lucene is a full-text search library which makes it easy to add search functionality to an application or website. It does so by adding content to a full-text index[1]. It then searches this index and returns results ranked by either the relevance to the query or by an arbitrary field such as a document's last modified date.

The advantage of Lucene is that it is able to achieve fast search responses because, instead of searching the text directly, it searches an index [2]. Searching requires an index to have already been built. It involves creating a query (usually via a Query Parser) and handing this query to an Index Searcher, which returns a list of Hits. Figure 1 shows the main module of Lucene[3].

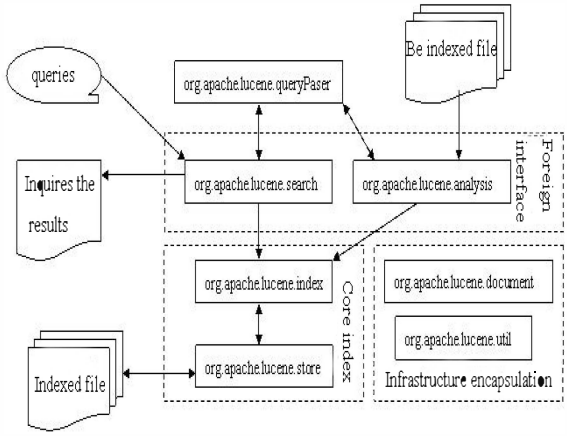

Fig.1 Lucene Module Diagram

From above Fig.1, it is clearly outlined that the Lucene consists of three parts: Foreign interface, Core index and Infrastructure encapsulation, of which the most important one is Core index part. Lucene is issued as Jar files, which is expressed as seven packages in Java. Each package has different functions, which is described in table [4]. In particular, the core packages are "org.apache.lucene.document","org.apache.lucene.analysis", "org.apache.lucene.index", and "org.apache.lucene.search"[5].

TABLE I. LUCENE PACKAGE STRUCTURE FUNCITON

package name	function discription
org.apache.lucene.queryPaser	query analyzer
org.apache.lucene.search	retr eval management
org.apache.lucene.analysis	language anaylyzer
org.apache.lucene.index	index management
org.apache.lucene.store	data store management
org.apache.lucene.document	indexed document streture management
org.apache.lucene.util	the util class provided for Lucene

III. SYSTEM DESIGN

The proposed online price comparison system PCS has designed and implemented a comparison system through the search engine. Basically, it is divided into four main modules: data crawling, data parsing, data indexing and information searching. The PCS system architecture is shown in Fig.2. The paper will illustrate the each module respectively as follows.

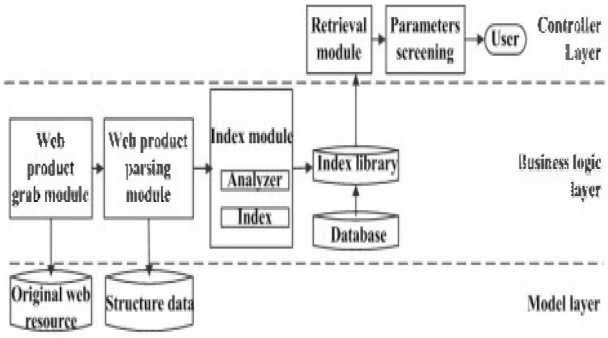

Fig.2 Price Comparison System Architecture

A. Data Crawling

The function of data crawling is to crawl the similar products information from various online stores according to the requirements of consumers. Generally, the process of data crawling is divided into three steps, the first step is to startup crawler for corresponding sites, then traverse and download the URL list which loaded in each crawler, the workflow is shown in Fig.3 as below:

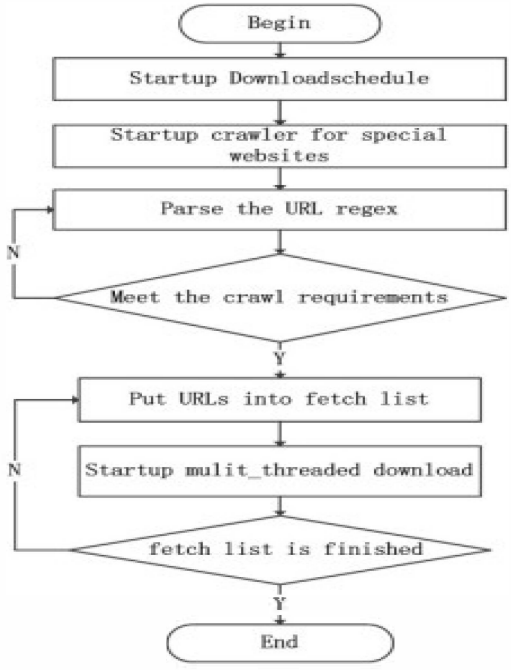

Fig.3 The Crawler Workflow

B. Data Parsing

The crawling data exists in the form of web pages and images, so that web data should be extracted，filt in order to get the relevant information. This format processing is called data parsing. Usually, the proposed system extract parameters from web pages and images then preserved them as a certain format. The workflow of data parsing is shown in Fig.4 as follows:

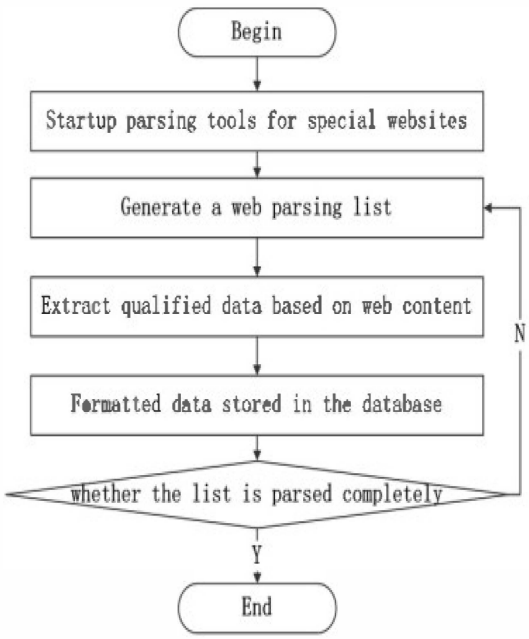

Fig.4 Data Parsing Workflow

C. Data Indexing

During the process of the data indexing, data have already been parsed out, we need to build indexing library for the coming information retrieval. Therefore, we select the open source Lucene to build index, the workflow of data indexing is shown in Fig.5.

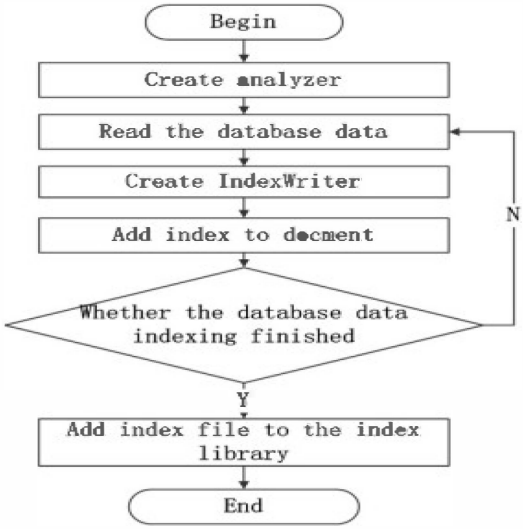

Fig.5 Indexing Data Workflow

D. Information Searching

The forth step is information searching, when users enter a keyword, the system will search the information in index library, then the search results will be presented in the form of a list of pages, the searching workflow is shown in Fig.6 as follows:

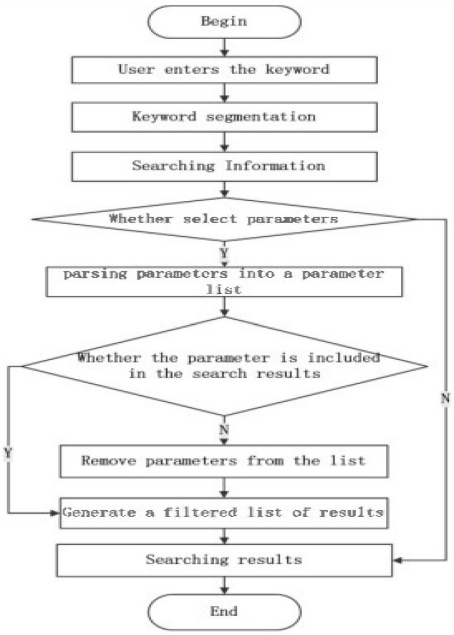

Fig.6 Information Searching Workflow

IV. SYSTEM IMPLEMENTATION

In the developing environment of the proposed system, the paper adopted web crawler, Lucene ,Jsoup ,SSI. The implementation of each module is described as follows.

A. Data Crawling

Due to the limited of server 503 error, the system cannot adopt Heritrix as the crawler. Therefore, the paper coded the new crawler following the Heritrix working principles, to grab the products information from the Amazon and the YHD websites precisely. The crawler's class package structure is shown in Fig.7 as follows:

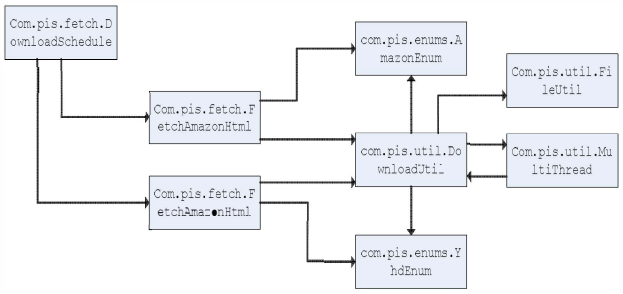

Fig.7 The Class Package Structure of Crawler

- DownloadSchedule: the class that is responsible for starting the back end crawl work from the seeds of the job;
- FetchAmazonHtml & FetchYhdHtm: the class of fetching pages and images to local disk;
- AmazonEnum, YhdEnum：they are the Amazon and YHD attributes enumerated classes;
- DownloadUtil: the core classes to parse the qualified URLs in the web;
- MultiThread: multi-threaded download tools;
- FileUtil: File operation class, used for file operation.

B. Data Parsing

The paper used Jsoup to parse crawling data, extracted data from web pages, then preserved attributes as a certain format into the database. The class structure of data parsing is shown in Fig.8 as follows:

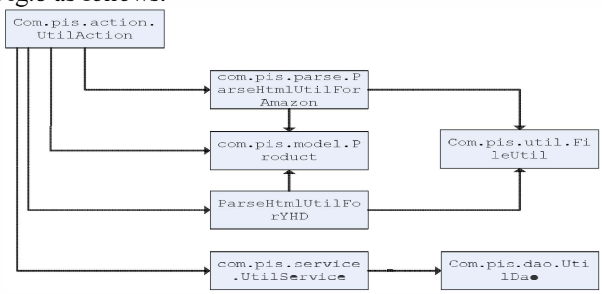

Fig.8 The Class Structure of Data Parsing

- UtilAction： The control layer of data parsing and indexing;
- ParseHtmlUtilForAmazon, ParseHtmlUtilForYHD：parsing tools of web pages for Amazon and YHD, mainly use Jsoup;
- UtilService：mainly responsible for the storage and indexing of format data;
- UtilDao： the class to call the configuration file sqlmap-product.xml of ibatis to achieve database CRUD operations , storage and read of indexing data.
- Product: the corresponding entity class of product database table.

C. Data Indexing

Data indexing is read out the database formatted data, the PCS calls the API of Lucene. In addition, due to the lack of Lucene for Chinese language search support, the proposed system combined with the Chinese sub-word Paoding software to improve the performance of Chinese character segmentation greatly, which provides better services for the information edited with Chinese languages. The critical code is shown in Table 2.

Table 2 The Code of Calling Paoding &Lucene

```
{……
/*
  PaodingAnalyzer() is called from another
  module that deal with the Chinese segmentation
*/
Analyzer analyzer = new PaodingAnalyzer();
Directory directory = FSDirectory.open(file);
IndexWriter indexWriter = new IndexWriter(directory,
analyzer,true,IndexWriter.MaxFieldLength.UNLIMITED);
indexWriter.deleteAll();
    for(Product product:list){
        doc = new Document();
        Field name = new Field(NAME,product.getName(),
            Field.Store.YES,Field.Index.NOT_ANALYZED);
        Field price = new Field(PRICE,product.getPrice(),
            Field.Store.YES,Field.Index.NOT_ANALYZED);
        doc.add(name);
        doc.add(price);
        indexWriter.addDocument(doc);
    }
    /*
     indexWriter is called
     from the Lucene module
    */
    indexWriter.optimize();
……
}
```

D. *Data Retrieval*

The search implementation uses Lucene to search and search class package structure shown in Fig.9.

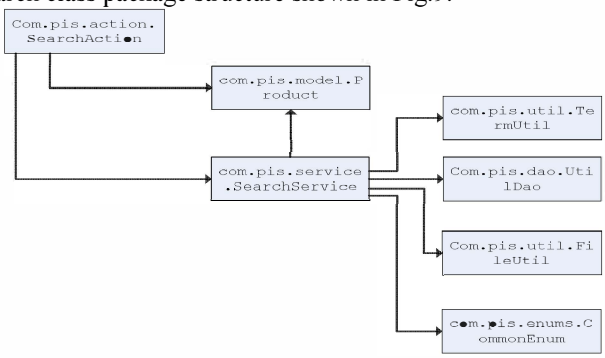

Fig.9. The Class Structure of Data Retrieval

- SearchAction： the control layer of search and keyword association.
- SearchService： mainly for the storage business processing of search and input keyword association.
- CommonEnum： store the index path and static corresponding properties constants of product, easy for system migration.
- TermUtil: tools operate on the index files, reading of the index data for keywords association.

V. EXPERIMENTAL RESULT AND ANALYSIS

Take an example of searching laptop information online, the PCS system will show the search results of laptops in Fig.11 as follow.

Fig. 11 The Laptop Information of Online

After clicking the product the user want to learn more information ,like an Apple laptop, the PCS will show the detailed information shown in the fig.12.

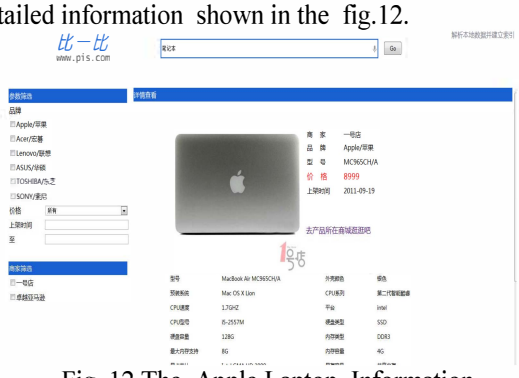

Fig. 12 The Apple Laptop Information

In addition, the PCS can supply searching results according to various parameters, for example , not only prices of product but also the products information on the same shop. If users want to learn all products in the No.1 shop, click the name icon of it, the result will shown in Fig.13 as follow.

Fig.13 The Products of All Brands In the Same Shop

VI. CONCLUSION

The traditional product search engines essentially provide basic browsing via category and “advanced” keyword without any change since the inception of online shopping[6][7]. The paper presents a product price comparison oriented product search engine, in order to show all the possible prices of products for customers.

The paper solves the problem efficiently, imitates Heritrix to write a multithreaded crawler in Java, then use Lucene and Mysql to develop a search system. In the end, the proposed system implemented web information crawling, data extraction and index building. The further work is to optimize individual requirements function in order to develop a more personalized search engine.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation (61075059), and Teaching Project of Hubei University of Technology (201106).We would like to thank all of people those assisted us during the process of developing the system.

REFERENCES

[1] Lucene Open Source Material.http://jakarta.apache.org/lucene

[2] Steven J. Owens, Lucene Tutorial,http://darksleep.com/lucene/.2008.7

[3] Jianxia Chen,Wei Wu,Ri Huang, A Mobile Phone Information Search Engine Based on Heritrix and Lucene, The 7th ICCSE, 2012.4

[4] Chun Liu, Analysis and Research of Web Chinese Retrieval System Based Lunece. 2009 First International Workshop on Education Technology and Computer Science.

[5] Su-TanYing, Guo-XianYong, Jin-Xin. One Chinese Full-Text Search System Based on Lucene. Computer Engineering, 2007.11,Volume3(23):94-96.

[6] Smith M D;Brynjolfsson E Consumer decision-making at an internet shopbot:Brand still matters 2001(04) doi:10.1111/1467-6451.00162

[7] Wu Da-zhong;Ray G;Geng Xian-jun Implications of reduced search cost and free riding in e-commerce 2004(02) doi:10.1287/mksc.1040.004