ATLANTIS PRESS
Advances in Computer Science Research (ACSR), volume 76
7th International Conference on Education, Management, Information and Mechanical Engineering (EMIM 2017)

KidSE: A Search Engine Designed for Children which Supports Simplified Chinese

Shulin Cao[1, a], Qingsong Lv[1, b], Yifan Wang[1, c], Qian Yin[1, d*] and Xin Zheng[1, e*]

[1]College of Information Science and Technology, Beijing Normal University, Beijing, China

[a]201511210110@mail.bnu.edu.cn, [b]201511210102@mail.bnu.edu.cn,
[c]201511210101@mail.bnu.edu.cn, [d]yinqian@bnu.edu.cn, [e]zhengxin@bnu.edu.cn

* The corresponding author

Keywords: Simplified Chinese; Children search engine; Lucene; Webpage filtering

Abstract. With the exploration of Internet information, searching has become an everyday task for both adults and children. Unlike adults, children have more difficulty identifying the Web information and they are more susceptible to unhealthy or misleading information. However, the existing search engines supporting simplified Chinese scarcely tackle this problem. In this paper, we propose a novel search engine called KidSE (kids' search engine) aiming at helping children to access Web knowledge safely and easily. First, the crawler collects information in children's websites. Then, the full-text inverted indexing mechanism and searching mechanism are designed by Lucene. In the process of information preprocessing and query filtering, unhealthy or misleading information is filtered effectively. Finally, Socket network programming is employed to achieve user-server interaction. Experiment shows that KidSE is effective and much safer than other children search engines such as "Baidu Children Search". The proposed work makes contributions to children search engines supporting simplified Chinese, which will have profound effects on children's health and growth.

Introduction

Internet access for children continues to increase over time and more and more children are utilizing search engine to get access to web information. Nevertheless, due to some characteristics such as relatively low cognition level and lack of internet experience, children encounter barriers in identifying web information and are more vulnerable to unhealthy or misleading information. Therefore, a need for a search engine catering for children arose.

While several influential children search engines supporting English are widely used, the search engines that support simplified Chinese are lagging behind. For example, "Baidu Children Search" once reported to be children-safe turns out to be no safer than general search engine Baidu.

In this paper, we develop a Lucene-based [1] children search engine called KidSE with the goal of assisting children in searching safely and easily. The safety is ensured from two respects. Firstly, when preprocessing information crawled from web, we utilize special algorithms to exclude the high-risk Webpages from index library. Thus, in principle, the unhealthy Webpages will not be indexed. Secondly, when parsing queries, we employ the IKAnalyzer [2] to filter the sensitive words (such as obscenities) based on the customized extension dictionary. High-risk information can be effectively filtered combining the two methods.

The remainder of this paper is organized as follows. First, the architecture of the system is presented. Then, the building of data warehouse module is introduced. Following this, the Indexing module is proposed. Then the Searching module is shown, which highlights the analyzing and filtering process, describes the details of advanced queries, and introduce the highlighting function. After this, the web interaction is established. Finally, experiment conducted, we conclude the paper and discuss future work.

Architecture of the System

The architecture of KidSE is depicted in Fig. 1.

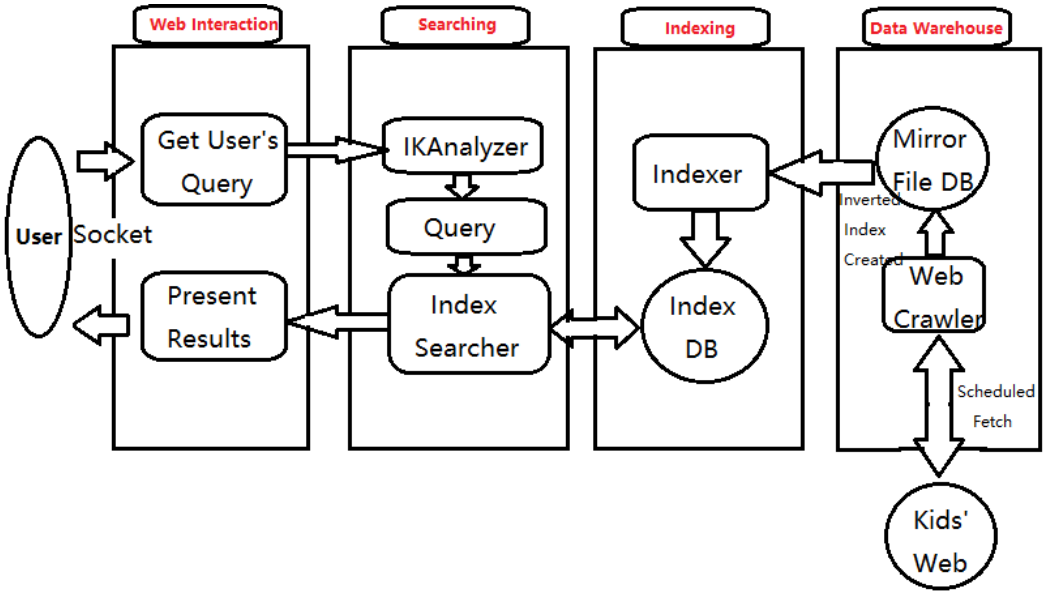

Figure 1. The architecture of the KidSE

Building Data Warehouse

Information Collection

Selecting seed URL. Since the search engine is oriented to children, based on surveys, we carefully select 30 high-quantity seed URL as the starting point the crawler traverses the network, which are the primary resources that children often use. Table 1 below shows part of the URL.

Traversing and Archiving. According to the 30 seed URL, web crawler get the page from web and analyze it, to extract all of the URL links and add them to the page data URL queue to be accessed, at the same time move the visited URL to the visited URL queue. All the collected pages are saved to store for further processing [3].The web crawler continuously repeat the above procedure, until the URL queue is empty. This run collects 57 GB and stores them in a local server.

Information Preprocessing

Extracting the Index Entry. After the information collection, we need to extract the preserved information and save them as the index entry.

Filtering High-risk Webpages. In the meantime, we employ the classic algorithm DFA [4] to exclude the high-risk webpages from index library to ensure the safety of children.

Table 1 Primary Children's Websites

http://child.iqiyi.com/	http://www.xiexingcun.com/	http://www.dm5u.com/
http://www.tigu.cn/	http://www.ppzuowen.com/	http://www.xj5u.com/

Indexing

Analyzing.

Chinese Word Segmentation. Different from English NLP [5], Chinese NLP systems need to begin with word segmentation, an unnecessary step in its English counterpart. The analysis rule of Lucene to English is simple. English takes space as a sign of the word segmentation. There is no space between the words of Chinese, and in the process of segmentation, we must encounter the ambiguity problems and unknown word identification, so the word segmentation for Chinese is much more difficult. The need for word segmentation creates additional challenges [6] for Analyzer. Here, we employ SmartChineseAnalyzer [7] which is a Lucene analysis package.

Chinese Character Encoding. To solve the problem of Chinese character encoding, we convert all existing texts to UTF-8 while the information preprocessing. In particular, we need the code below when indexing.

Document. add(new TextField("content", new BufferedReader(new InputStreamReader(stream, StandardCharsets.UTF_8))));

Lucene index package. We employ Lucene, which uses an inverted index technique in managing and creating indexes over a collection of documents [8].

Two index class. There are two index classes that can be used.

IndexWebpages. Index Webpages handles the information preprocessed as is introduced before. The structure of documents is depicted in the Table 2 as follows. In particular, the field "main domain" is aimed to support advanced site-specific query which is to be discussed in the following module.

IndexBooks. IndexBooks indexes all the books (PDF link) for children.

Table 2 Structure of Document

	Abstract(sabstract)	image	link	main domain	title
type	text(string)	string	string	string	string

Searching

Analyzing and Filtering. IKAnalyzer is a light-weight java-based third party kit of the Chinese word segmentation. The IKtokenizer applies the "positive iteration most fine-grained segmentation algorithm" and the multiprocessing model, which supports English, figures and Chinese characters. In particular, it supports custom dictionary that is of great help to filtering the high-risk query terms. Hence, we employ the IKAnalyzer, customize the sensitive word extension dictionary and configure the IKAnalyzer.cfg.xml. To the aim of filtering the unhealthy and misleading words for children, we add the customized dictionaries shown in Fig. 2.

Advanced Query.

Site-specific Query. As depicted in table 2, the field "main domain" has been added to the index entry, which contains the main domain of corresponding link. For example, to search all the "song" of "http://child.iqiyi.com/", we can input "site: child.iqiyi.com song". To implement site-specific query, we employ BooleanQuery which matches documents matching boolean combinations of other queries. In particular, the method add (query, BooleanClause. Occur. FILTER) is utilized, whose operator FILTER means that the query must appear in the matching documents without participating in scoring. The core code is as follows:

```
Query queryabstract=new TermQuery(new Term("abstract", s[1]));
Query querysite=new TermQuery(new Term("site",ss[1]));
BooleanQuery query=new
BooleanQuery.Builder().add(querysite,BooleanClause.Occur.FILTER).add(queryabstract,BooleanC
lause.Occur.SHOULD).build();
```

Wildcard Query. Using the package wildquery [9]. The core code is as follows:

```
Query query = new WildcardQuery(new Term("abstract", keywords));
```

Fuzzy Query. Using the package fuzzyquery [10]. The core code is as follows:

```
Query query = new FuzzyQuery(new Term("abstract", keywords));
```

Highlighting. Highlighting is the ability to highlight the text that match the query in the information of the results displayed to the users [11]. The highlight package contains classes to provide "keyword in context" features typically used to highlight search terms in the text of results pages.

Web Interaction

In order to enable children to obtain web content from the server, we utilize the Socket programming to establish a connection between the client and server. The procedure is shown in Figure 3.

Server. As for the server, we employ Socket in java. The core code is as follows.

```
ServerSocket server=new ServerSocket(5678);
Socket client=server.accept();
BufferedReader in=new BufferedReader(new InputStreamReader (client. getInputStream()));
PrintWriter out=new PrintWriter (client.getOutputStream());
```

Client. As for the client, we employ Socket in PHP. The core code is as follows:

```
$socket = socket_create(AF_INET, SOCK_STREAM, SOL_TCP);
$result = socket_connect($socket, $ip, $port);
socket_write($socket, $in, strlen($in);
while($out = socket_read($socket,1024)) {
        echo mb_convert_encoding($out, "utf-8", "gb2312");
}
```

Figure 2. The configuration of dictionaries

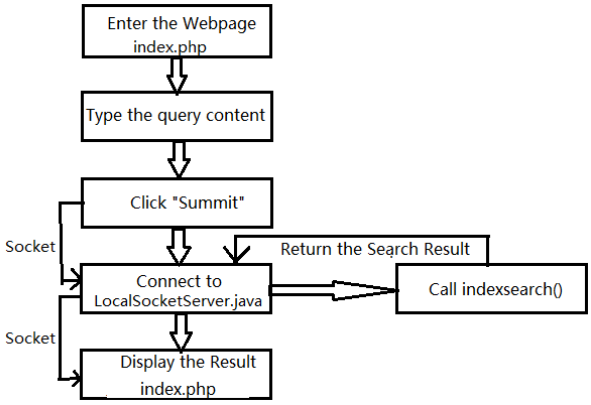

Figure 3.The procedure of web interaction

Summary

In this paper, we proposed the KidSE, a novel children search engine which supports simplified Chinese. Designed to assist children in accessing Web information safely and easily, the Lucene-based KidSE consists of a Web crawler, an Indexer, a Searcher and a Web interface. First, the crawler collects information in children's sites such as "kid.qq.com" which are primary resources of children. Then, the full-text inverted indexing mechanism is designed by Lucene. Following this, we implement the searching module, including advanced query. In the process of information preprocessing and query filtering, unhealthy or misleading information is filtered effectively, ensuring the safety and health of children. In the last section, Socket network programming is employed to achieve user-server interaction. To evaluate the system, we conducted a case study with 30 children (aged between 8 -13). Experiment shows that KidSE is effective and much safer than other children search engines such as "Baidu Children Search". We anticipate that KidSE can have a positive effect on children's growth.

References

[1] Information on http://lucene.apache.org/
[2] Information on https://code.google.com/archive/p/ik-analyzer/ (in Chinese)

[3] Hai Zhao, C.N. Huang. Effective tag set selection in Chinese word Segmentation via conditional random field modeling [C], In: Proceedings of PA-CL IC220.Wu Han, November 123, 2006: 84-94.
[4] Information on https://en.wikipedia.org/wiki/Natural_language_processing
[5] Zhang S, Kang T, Zhang X, et al. Speculation detection for Chinese clinical notes: Impacts of word segmentation and embedding models[J]. Journal of biomedical informatics, 2016, 60: 334-341.
[6] Information on http://lucene.apache.org/core/6_2_1/analyzers-smartcn/index.html
[7] Lee H B, Nazareno F, Jung S H, et al. A vertical search engine for school information based on heritrix and lucene [C]//International Conference on Hybrid Information Technology. Springer Berlin Heidelberg, 2011: 344-351.
[8] Alani H, Kim S, Millard D E, et al. Automatic ontology-based knowledge extraction from web documents [J]. IEEE Intelligent Systems, 2003, 18(1): 14-21.
[9] Information on http://lucene.apache.org/core/6_2_1/core/index.html
[10] Information on http://lucene.apache.org/core/6_2_1/core/index.html
[11] Armentano M G, Godoy D, Campo M, et al. NLP-based faceted search: Experience in the development of a science and technology search engine [J]. Expert Systems with Applications, 2014, 41(6): 2886-2896.