ISSN 2320 – 2602
Volume 4 No.4, April 2015
International Journal of Advances in Computer Science and Technology
Available Online at http://www.warse.org/ijacst/static/pdf/file/ijacst12442015.pdf

Implementation of a Web Search Engine for Restaurants using Lucene

Sujoung Oh[1], Minsoo Lee[2]
[1]Computer Science and Engineering, Ewha Womans University, Korea, crystal7862@gmail.com
[2]Computer Science and Engineering, Ewha Womans University, Korea, mlee@ewha.ac.kr

ABSTRACT

Recently, many people use Internet blogs or Website related to famous restaurants. People are interested in information found from various sources and want to efficiently search them. These search engines should be easy to develop and effectively enhanced using open source search engines to provide for various international language specific techniques to be experimented for search. Therefore, we developed a search engine to find popular restaurants based on Lucene and used information retrieval techniques. The proposed search engine can extract restaurant lists related to keywords. The search engine, also, provides functions to be redirected to the original restaurant website to directly verify the information found.

Key words : Lucene, search engine, restaurant, information retrieval

1. INTRODUCTION

Many people visit websites providing restaurant information to find a restaurant which serves delicious food. Their search keywords are usually the menu or the location. As such there is a need for various search mechanisms for various user inputs on various data for restaurants which is a very popular topic on the web. However search engines have limited capability in terms of the data that they use or the queries that they can support. Thus various techniques need to be experimented. This is also true in an international environment where different requirements exist for specific languages and countries. These search engines should be easy to develop and effectively enhanced using open source search engines to provide for various international language specific techniques to be experimented for search. For this reason we built a general web search engine using Lucene. And if a user wants to know more information about the restaurant, the user can connect to the website.

This paper is organized as follows. Section 2 introduces related works and section 3 describes the methodology proposed in this paper and the implementation results. And finally in section 4 we describe the conclusions and future work.

2. RELATED WORKS

2.1 Lucene

Apache Lucene is a free and open source information retrieval software library, originally created in Java by Doug Cutting. It is supported by Apache Software Foundation and is released under the Apache Software License. Lucene has been ported to other programming languages including Delphi, Perl, C#, C++, Python, Ruby, and PHP. While suitable for any application which requires full text indexing and searching capability, Lucene has been widely recognized for its utility in the implementation of Internet search engines and local, single-site searching. At the core of Lucene’s logical architecture is the idea of a document containing fields of text. This flexibility allows Lucene’s API to be independent of the file format. Text from PDFs, HTML, Microsoft Word, and OpenDocument documents, as well as many others (except images), can all be indexed as long as their textual information can be extracted[1,2].

2.2 Web Search Engines

A web search engine is a software system for searching information on the World Wide Web. The search results are presented in a line of results often referred to as SERPs(Search Engine result pages). The information can be web pages, images, and other type of files. And search engines can be used for mining data in databases[3].

A search engine operates via the following three steps. The first step is Web crawling. The second step is indexing. The third step is searching. A web crawler is an Internet bot that browses the World Wide Web, typically for the purpose of Web indexing[4]. Search engine indexing collects, parses, and stores data to facilitate fast and accurate information retrieval. Index design incorporates concepts from linguistics, cognitive psychology, mathematics, informatics, and computer science. An alternate name for the process in the context of search engines designed to find web pages on the Internet is web indexing[5]. In the search step, a web search query that a user enters into a web search engine to satisfy his or her information needs is issued. The web search queries are distinctive in that they are often plain text or hypertext with optional search-directives[6].

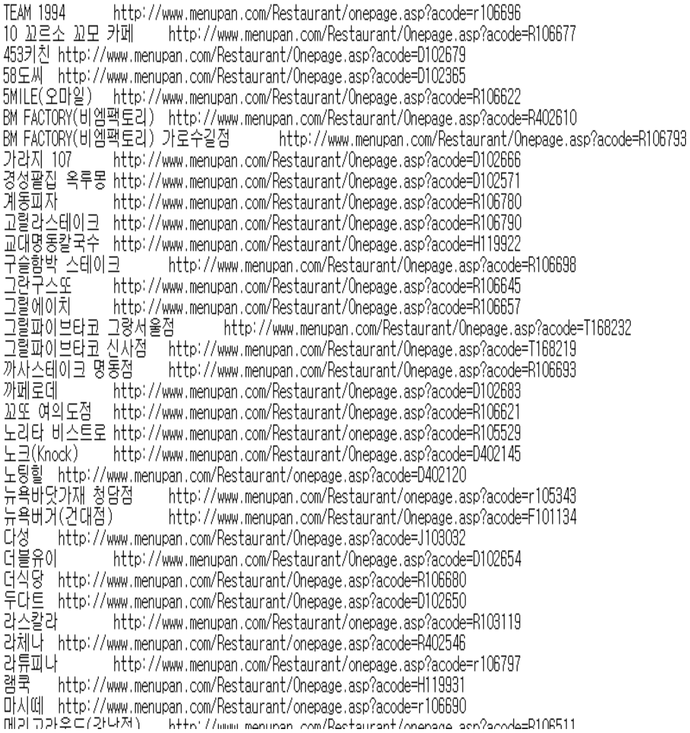
[1] This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(No. 2012R1A1A2006850).

3. SYSTEM DEVELOPMENT

The system development is divided into three parts. The first part is about extracting the text from the Web site URL. The second part is about creating the index file using the information of the web site. And finally, we describe the searching for the results.

3.1 Extracting Text using URLs

We extract the text from web documents using the URL information. When we save the text file with the needed information, the format we use is the Json format to enable Json parsing.



```
TEAM 1994       http://www.menupan.com/Restaurant/onepage.asp?acode=r106696
10 꼬르소 꼬모 카페      http://www.menupan.com/Restaurant/Onepage.asp?acode=R106677
453키친 http://www.menupan.com/Restaurant/Onepage.asp?acode=D102679
58도씨  http://www.menupan.com/Restaurant/Onepage.asp?acode=D102365
5MILE(오마일)   http://www.menupan.com/Restaurant/Onepage.asp?acode=R106622
BM FACTORY(비엠팩토리)  http://www.menupan.com/Restaurant/Onepage.asp?acode=R402610
BM FACTORY(비엠팩토리) 가로수길점       http://www.menupan.com/Restaurant/Onepage.asp?acode=R106793
가라지 107      http://www.menupan.com/Restaurant/Onepage.asp?acode=D102666
경성팥집 옥루몽 http://www.menupan.com/Restaurant/Onepage.asp?acode=D102571
계동피자        http://www.menupan.com/Restaurant/Onepage.asp?acode=R106780
고릴라스테이크  http://www.menupan.com/Restaurant/Onepage.asp?acode=R106790
교대명동칼국수  http://www.menupan.com/Restaurant/Onepage.asp?acode=H119922
구슬함박 스테이크       http://www.menupan.com/Restaurant/Onepage.asp?acode=R106698
그란구스또      http://www.menupan.com/Restaurant/Onepage.asp?acode=R106645
그릴에이치      http://www.menupan.com/Restaurant/Onepage.asp?acode=R106657
그릴파이브타코 그랑서울점       http://www.menupan.com/Restaurant/Onepage.asp?acode=T168232
그릴파이브타코 신사점   http://www.menupan.com/Restaurant/Onepage.asp?acode=T168219
까사스테이크 명동점     http://www.menupan.com/Restaurant/Onepage.asp?acode=R106693
까페로데        http://www.menupan.com/Restaurant/Onepage.asp?acode=D102683
꼬또 여의도점   http://www.menupan.com/Restaurant/Onepage.asp?acode=R106621
노리타 비스트로 http://www.menupan.com/Restaurant/onepage.asp?acode=R105529
노크(Knock)     http://www.menupan.com/Restaurant/Onepage.asp?acode=D402145
노팅힐  http://www.menupan.com/Restaurant/onepage.asp?acode=D402120
뉴욕바닷가재 청담점     http://www.menupan.com/Restaurant/onepage.asp?acode=r105343
뉴욕버거(건대점)        http://www.menupan.com/Restaurant/Onepage.asp?acode=F101134
다성    http://www.menupan.com/Restaurant/Onepage.asp?acode=J103032
더블유이        http://www.menupan.com/Restaurant/Onepage.asp?acode=D102654
더식당  http://www.menupan.com/Restaurant/Onepage.asp?acode=R106680
두다트  http://www.menupan.com/Restaurant/Onepage.asp?acode=D102650
라스칼라        http://www.menupan.com/Restaurant/onepage.asp?acode=R103119
라체나  http://www.menupan.com/Restaurant/onepage.asp?acode=R402546
라튜피나        http://www.menupan.com/Restaurant/onepage.asp?acode=r106797
램쿡    http://www.menupan.com/Restaurant/Onepage.asp?acode=H119931
마시떼  http://www.menupan.com/Restaurant/onepage.asp?acode=r106690
메리고라운드(강남점)    http://www.menupan.com/Restaurant/onepage.asp?acode=R106511
```

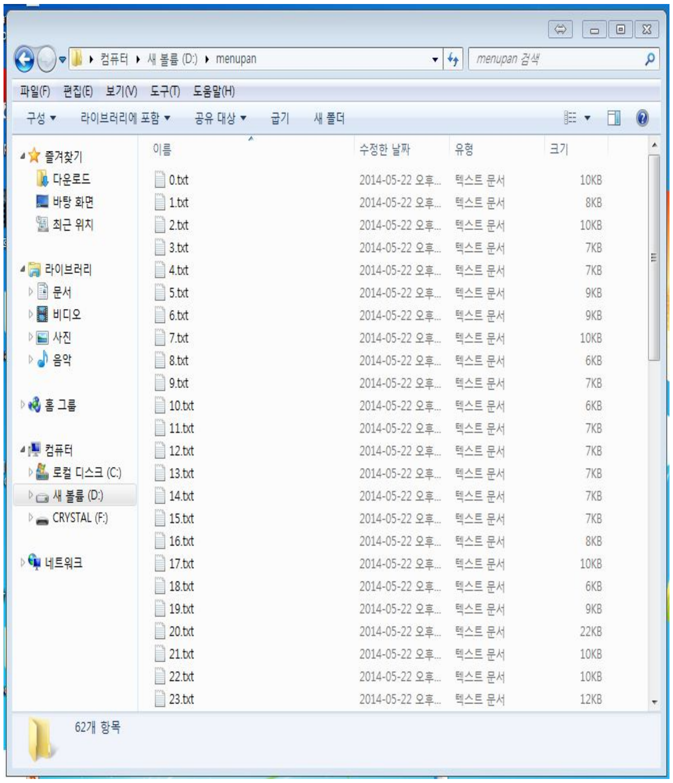
Figure 1: Restaurant name & URL

Figure 1 illustrates the information that contains the restaurant name and the URL of the web site. The Web crawler can continuously traverse among the URLs to create a larger subset of relevant websites. From the web sites, we can extract restaurant data such as menu, location, and client comment.

Figure 2 presents the list that contains the restaurant data. Each text file indicates the data for one restaurant.

3.2 Index Creation

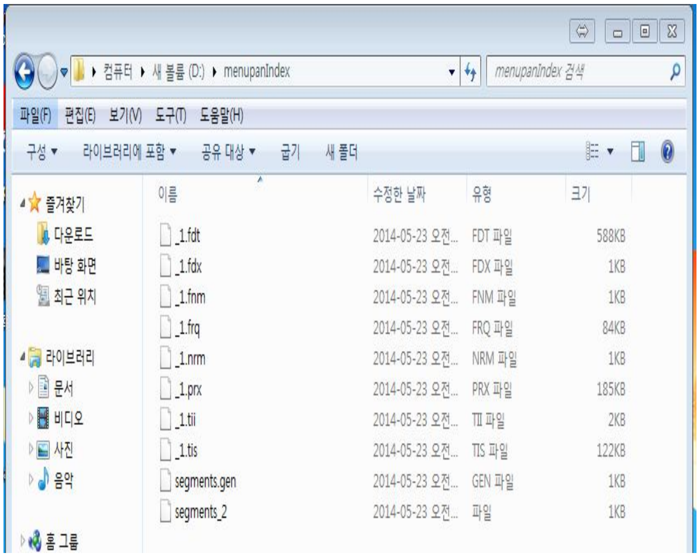
We read the text files, and then create the index for the restaurant data. It contains the restaurant name, main text, and URL. Figure 3 illustrates the files in the index folder. Lucene is used for the creation of index structures.



이름	수정한 날짜	유형	크기
0.txt	2014-05-22 오후...	텍스트 문서	10KB
1.txt	2014-05-22 오후...	텍스트 문서	8KB
2.txt	2014-05-22 오후...	텍스트 문서	10KB
3.txt	2014-05-22 오후...	텍스트 문서	7KB
4.txt	2014-05-22 오후...	텍스트 문서	7KB
5.txt	2014-05-22 오후...	텍스트 문서	9KB
6.txt	2014-05-22 오후...	텍스트 문서	9KB
7.txt	2014-05-22 오후...	텍스트 문서	10KB
8.txt	2014-05-22 오후...	텍스트 문서	6KB
9.txt	2014-05-22 오후...	텍스트 문서	7KB
10.txt	2014-05-22 오후...	텍스트 문서	6KB
11.txt	2014-05-22 오후...	텍스트 문서	7KB
12.txt	2014-05-22 오후...	텍스트 문서	7KB
13.txt	2014-05-22 오후...	텍스트 문서	7KB
14.txt	2014-05-22 오후...	텍스트 문서	7KB
15.txt	2014-05-22 오후...	텍스트 문서	7KB
16.txt	2014-05-22 오후...	텍스트 문서	8KB
17.txt	2014-05-22 오후...	텍스트 문서	10KB
18.txt	2014-05-22 오후...	텍스트 문서	6KB
19.txt	2014-05-22 오후...	텍스트 문서	9KB
20.txt	2014-05-22 오후...	텍스트 문서	22KB
21.txt	2014-05-22 오후...	텍스트 문서	10KB
22.txt	2014-05-22 오후...	텍스트 문서	10KB
23.txt	2014-05-22 오후...	텍스트 문서	12KB

62개 항목

Figure 2: Extracted restaurant data files



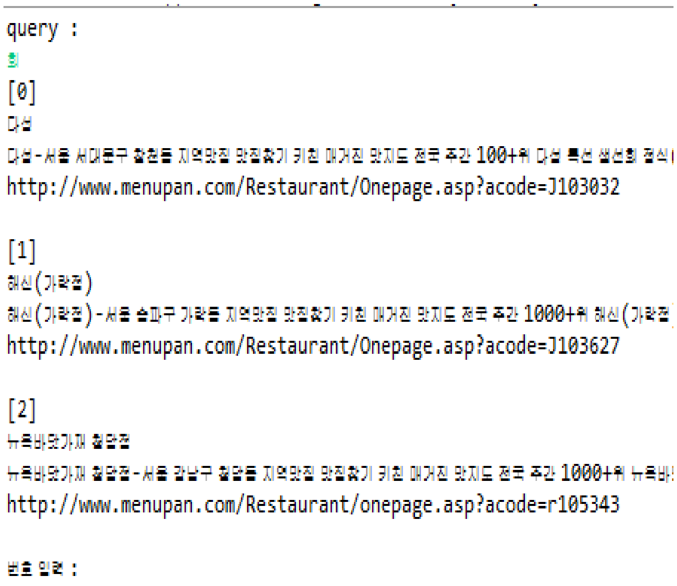
컴퓨터 ▸ 새 볼륨 (D:) ▸ menupanIndex
menupanIndex 검색
파일(F) 편집(E) 보기(V) 도구(T) 도움말(H)
구성 ▾ 라이브러리에 포함 ▾ 공유 대상 ▾ 굽기 새 폴더
즐겨찾기
다운로드
바탕 화면
최근 위치
라이브러리
문서
비디오
사진
음악
홈 그룹

이름	수정한 날짜	유형	크기
_1.fdt	2014-05-23 오전...	FDT 파일	588KB
_1.fdx	2014-05-23 오전...	FDX 파일	1KB
_1.fnm	2014-05-23 오전...	FNM 파일	1KB
_1.frq	2014-05-23 오전...	FRQ 파일	84KB
_1.nrm	2014-05-23 오전...	NRM 파일	1KB
_1.prx	2014-05-23 오전...	PRX 파일	185KB
_1.tii	2014-05-23 오전...	TII 파일	2KB
_1.tis	2014-05-23 오전...	TIS 파일	122KB
segments.gen	2014-05-23 오전...	GEN 파일	1KB
segments_2	2014-05-23 오전...	파일	1KB

Figure 3: Index file creation

3.3 Search

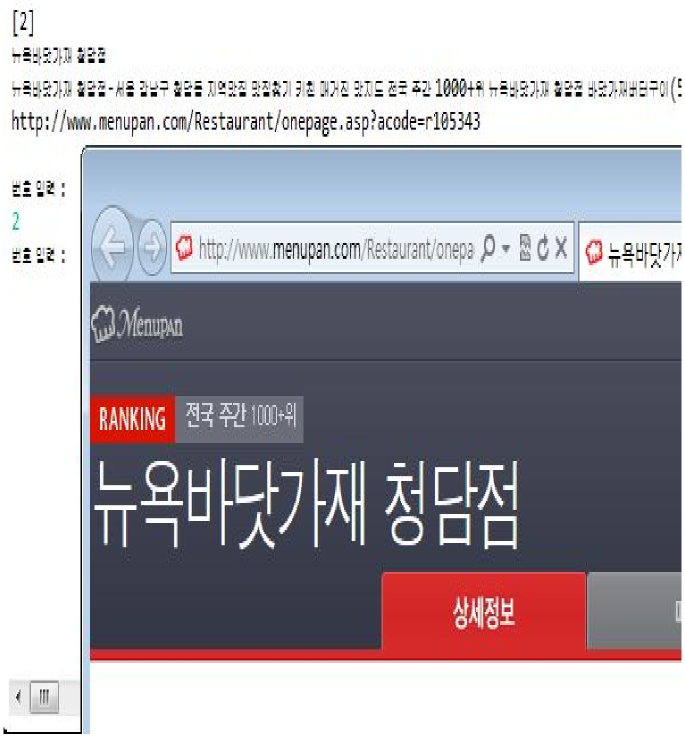
Our engine is based on Lucene and will return the search results based on the keyword. The results include the input query, restaurant lists about the input query, address of the restaurant, and URL of the restaurant.



```
query :
회
[0]
다설
다설-서울 서대문구 창천동 지역맛집 맛집찾기 키친 매거진 맛지도 전국 주간 100+위 다설 특선 생선회 정식(
http://www.menupan.com/Restaurant/Onepage.asp?acode=J103032

[1]
해신(가락점)
해신(가락점)-서울 송파구 가락동 지역맛집 맛집찾기 키친 매거진 맛지도 전국 주간 1000+위 해신(가락점
http://www.menupan.com/Restaurant/Onepage.asp?acode=J103627

[2]
뉴욕바닷가재 청담점
뉴욕바닷가재 청담점-서울 강남구 청담동 지역맛집 맛집찾기 키친 매거진 맛지도 전국 주간 1000+위 뉴욕바
http://www.menupan.com/Restaurant/onepage.asp?acode=r105343

번호 입력 :
```

Figure 4: Search results

Figure 4 shows the search results. In this example, we used 'sushi' as the keyword. After searching, we got three restaurant data related to sushi.



```
[2]
뉴욕바닷가재 청담점
뉴욕바닷가재 청담점-서울 강남구 청담동 지역맛집 맛집찾기 키친 매거진 맛지도 전국 주간 1000+위 뉴욕바닷가재 청담점 바닷가재바비큐이(
http://www.menupan.com/Restaurant/onepage.asp?acode=r105343

번호 입력 :
2
번호 입력 :
```

Figure 5: Connecting to the Web site

And our engine can connect to the web site of the restaurant when the information on the web site needs to be verified. The web site is provided when a user selects the restaurant index number. Figure 5 shows the restaurant web site which is accessed through the developed search engine.

4. CONCLUSION & FUTURE WORK

This paper only search the documents contained search keyword. For this reason, our engine is hard to find popular restaurant. To solve the problem, in the future work, we develop search engine that use ranking by adding the heuristic values such as TV program (that recommend popular restaurant) and positive word such as delicious, excellent, wonderful, and "visit again".

REFERENCES

1. Lucene, Wikipedia, [Online] Available: http://en.wikipedia.org/wiki/Lucene
2. M. McCandless, E. Hatcher, O. Gospodnetic, Lucene IN ACTION, 2nd ed., Manning Publications, 2010.
3. Web Search Engine, Wikipedia, [Online] Available: http://en.wikipedia.org/wiki/Web_search_engine
4. Web Crawler, Wikipedia, [Online] Available: http://en.wikipedia.org/wiki/Web_crawler
5. Indexing, Wikipedia, [Online] Available: http://en.wikipedia.org/wiki/Searcg_engine_indexing
6. Web Search Query, Wikipedia, [Online] Available: http://en.wikipedia.org/wiki/Web_search_query