OU Web Search Engine based on Apache Nutch, Solr, and Lucene

Xin Ye xy348709@ohio.edu

Click the link below to use the search engine:

http://xinye-ohiou.rhcloud.com/cs6900/index.html

1. Crawl 117,620 pages using Nutch

(1). Install and configure Nutch

a. Download the source code of Apache Nutch 1.7 for this project

b. Apply a patch for fixing bug 1640 of Nutch 1.7

```
patch –p0 < NUTCH-1640.patch
```

c. Compile

```
ant
```

d. export JAVA_HOME=/System/Library/Frameworks/JavaVM.framework/Versions/1.6/Home

e. Add following information in nutch-site.html

```
<!--Crawler agent name -->
<property>
      <name>http.agent.name</name>
      <value>Macbook Air Home</value>
</property>
<!-Set delay to be 100ms -->
<property>
      <name>fetcher.server.delay</name>
      <value>0.1</value>
</property>
<!--Use 200 threads for crawling -->
<property>
      <name>fetcher.threads.fetch</name>
      <value>200</value>
</property>
```

f. Create *urls/seed.txt* with 10 seeds including

```
http://www.ohio.edu/
http://www.library.ohiou.edu/
http://www.ohio.edu/athens/bldgs/arc.html
http://www.ohio.edu/healthalerts
http://www.ohioalumni.org/give-to-ohio
http://www.ohio.edu/disabilities
http://www.facilities.ohiou.edu/cats
```

http://www.ohio.edu/compass
http://www.ohio.edu/hr
http://www.ohio.edu/admissionsg

g. Edit the regular expression in *regex-urlfilter.txt* to match only ohio/ohiou.edu domain

```
# accept anything else
+^http://([a-z0-9]*\.)*ohio(u?).edu/
```

(2). Manually crawl more than 117K pages

h. Inject seeds

```
bin/nutch inject crawl/crawldb urls
```

i. Generate a list of pages that to be crawled

```
bin/nutch generate crawl/crawldb crawl/segments
```

j. Fetch pages

```
bin/nutch fetch crawl/segments/segment1
```

k. Parse pages

```
bin/nutch parse crawl/segments/segment1
```

l. Update database

```
bin/nutch updatedb crawl/crawldb crawl/segments/segment1
```

m. Check how many pages have been crawled now

```
bin/nutch readdb crawl/crawldb –stats
```

n. Repeat step i to m above to continue to crawl pages until more than 117K pages have been crawled.

o. Invert all links for further indexing anchor text with the pages that have been crawled.

```
bin/nutch invertlinks crawl/linkdb -dir crawl/segments
```

2. Index using Nutch and Solr

(1). Install and configure Solr

a. Download and unzip the binary file of Solr 4.6 for this project

b. Copy *apache-nutch-1.7/runtime/local/conf/schema-solr4.xml* from Nutch 1.7 to Solr 4.6 *solr-4.6.0/example/solr/new_core/conf/schema.xml*

c. Edit *schema.xml* to store content information of pages for further highlighting query terms in displayed snippets

```
<field name="content" type="text_general" stored="true" indexed="true"/>
```

d. Add in line 351 of *schema.xml* of the below missing version information

```
<field name="_version_" type="long" indexed="true" stored="true"/>
```

e. Run Solr under *solr-4.6.0/example*

```
java -jar start.jar
```

f. Open http://localhost:8080/solr to manage Solr

h. To add a new core, simply click “Core Admin”->”Add Core”, copy *solr-4.6.0/example/solr/old_core/conf to solr-4.6.0/example/solr/new_core/conf,* and restart Solr again.

(2). Manually index crawled pages into Solr

i. Index pages in one segment into Solr

```
bin/nutch solrindex http://localhost:8983/solr/new_core crawl/crawldb/ -linkdb crawl/linkdb/ crawl/segments/segment1/ -filter -normalize
```

j. Repeat step i above until pages in all segments have been indexed into Solr

k. Delete duplicates

/bin/nutch solrdedup http://localhost:8983/solr/new_core

l. Clean Solr by deleting gone pages

/bin/nutch solrclean crawl/crawldb/ http://localhost:8983/solr/new_core

m. Optimize Solr index by clicking “Optimized” in http://localhost:8080/solr the new_core “Overview” page

Now we have used Nutch 1.7 to crawl 117,620 pages, have integrated Solr 4.6 with Nutch 1.7, and have indexed all the crawled pages into Solr 4.6. The index in Solr 4.6 can be used by Apache Lucene 4.6.0 (Note: the index created by Solr 4.6 can be used only by Lucene 4.6 but not Lucene 4.5.x or lower versions).

3. Search using Lucene

The Java servlet source code is TestServlet.java

The main html page is index.html

(1). Use Tomcat and Lucene in Eclipse

a. Download and unzip the binary file of Tomcat 7.0.47 for this project.

b. Add Tomcat server in Eclipse.

c. Download Lucene 4.6.0 for this project.

d. Create a new “Dynamic Web Project” in Eclipse, and create a new “Servlet” TestServlet.java in this project.

e. Put index.html and background.jpg under “WebContent”.

f. Copy necessary Lucene *.jar library files into /WebContent/WEB-INF/lib/

This project uses lucene-analyzers-common-4.6.0.jar, lucene-core-4.6.0.jar, lucene-highlighter-4.6.0.jar, and lucene-queryparser-4.6.0.jar, and add these jar files in the buildpath

(2). Deploy the servlet

g. Export the whole project under Eclipse to a WAR file *cs6900.war*

h. Copy the exported WAR file *cs6900.war* to *apache-tomcat-7.0.47/webapps/*

i. Run Tomcat, and use the search engine by http://ipaddress:8080/cs6900

4. Result Analysis

(1) Comparison with Google

Query	VSM(content) + 0.5 * VSM(anchor text)		Google (site:ohio.edu or site:ohiou.edu)	
	The relevant page in Top10	Precision @10	The relevant page in Top10	Precision @10
1. OU Home Page	http://www.ohio.edu/ (rank 1)	0.1	http://www.ohio.edu/ (rank 1)	0.1
2. OU Library	www.ohio.edu/athens/bldgs/alden.html (rank 4)	0.1	www.ohio.edu/athens/bldgs/alden.html (rank 1) All are relevant.	1
3. ARC building	http://www.ohio.edu/athens/bldgs/arc.html (rank 1)	0.1	http://www.ohio.edu/athens/bldgs/arc.html (rank 1)	0.1
4. Health Alerts	http://www.ohio.edu/alert/ (rank 7)	0.1	https://www.ohio.edu/healthalerts/ (rank 1) All are relevant.	1
5. Give to OHIO	http://www.ohio.edu/advancement/gift.cfm (rank 6)	0.1	http://www.ohio.edu/advancement/gift.cfm (rank 1) Other 7 are relevant.	0.8
6. ADA compliance	www.ohio.edu/disabilities/ (rank 1) http://www.ohio.edu/athens/ada_access.html (rank 3)	0.2	www.ohio.edu/disabilities/ (rank 2) Other 3 are relevant.	0.4
7. CATS	http://www.ohio.edu/sustainability/programs/transportation.cfm (rank 2) www.facilities.ohiou.edu/cats/ (rank 3) and other 3	0.5	www.facilities.ohiou.edu/cats/ (rank 1) and other 3	0.4
8. University News and Updates	http://www.ohio.edu/ (rank 1)	0.1	www.ohio.edu/compass (rank 1) and other 9	1
9. HR	http://www.ohio.edu/employees/ (rank 1) http://www.ohio.edu/hr/ (rank 7) and other 5	0.7	www.ohio.edu/hr and other 9	1
10. Admissions Web Site	http://www.ohio.edu/admissions/ (rank 3) and other 2	0.3	www.ohio.edu/admissions (rank 1) and other 7	0.8
11. Ohio University	http://www.ohio.edu/ (rank 1) and other 9	1	http://www.ohio.edu/ (rank 1) and other 9	1
12. Baker Center	http://www.ohio.edu/athens/parking/wgreen.html (rank 1) http://www.ohio.edu/athens/bldgs/newbaker.html (rank 9) and other 3	0.5	www.ohio.edu/eventservices/baker (rank 1) and other 8	0.9
13. EECS	www.ohio.edu/eecs (rank 1) and other 4	0.5	www.ohio.edu/eecs (rank 1) and all	1

14. Meal Plan	None.	0	http://www.ohio.edu/food/plans/residential.cfm (rank 1) and all	1
15. Football	http://www.ohio.edu/athens/bldgs/stadium.html (rank 5) http://www.ohio.edu/recreation/intramural/rules.cfm (rank 9)	0.2	http://www.sportsad.ohio.edu/events/bobcat-blackout/ (rank 2) and other 4	0.5
16. Calendar	http://www.ohio.edu/calendar/ (rank 1) and other 2	0.3	http://www.ohio.edu/education/news-and-events/calendar.cfm (rank 1) and other 8	0.9
17. Software Engineering	http://www.ohio.edu/athens/bldgs/stocker.html (rank 3) http://www.ohio.edu/engineering/index.cfm (rank 4) and other 3	0.5	http://engineering.online.ohio.edu/do-it-better/better-engineers/software-engineering/ (rank 1) and other 7	0.8
18. Student Expo	http://www.ohio.edu/graduate/studentexpo.cfm (rank 3) and other 3	0.4	http://www.ohio.edu/graduate/studentexpo.cfm (rank 1) and other 9	1
19. Police	www.ohio.edu/police (rank 1) and other 1	0.2	www.ohio.edu/police (rank 1) and other 9	1
20. Xin Ye	http://vital.cs.ohio.edu/ (rank 1)	0.1	http://vital.cs.ohio.edu/vitalwiki/index.php/User:XinYe (rank 1) and other 9	1

(2) Comparison among VSM, BM25, and LM

Query	Precision@10					
	Do not use anchor text			Similarity(content) + 0.5 * Similarity(anchor text)		
	VSM	BM25	LM	VSM	BM25	LM
1. OU Home Page	0	0	0	0.1	0.1	0.1
2. OU Library	0.1	0.1	0.3	0.1	0.1	0.3
3. ARC building	0.1	0.1	0.1	0.1	0.1	0.1
4. Health Alerts	0.1	0.1	0.7	0.1	0.1	0.7
5. Give to OHIO	0.1	0	0	0.1	0	0.1
6. ADA compliance	0	0	0	0.2	0	0.2
7. CATS	0.5	0.5	0	0.5	0.5	0
8. University News and Updates	0.1	0.1	0.1	0.1	0.1	0.1

9. HR	0.6	0.6	0.1	0.7	0.6	0.1
10. Admissions Web Site	0.2	0.3	0.2	0.3	0.3	0.4
11. Ohio University	1	1	0.3	1	1	0.3
12. Baker Center	0.5	0.5	0.5	0.5	0.5	0.5
13. EECS	0.6	0.6	0	0.5	0.6	0
14. Meal Plan	0	0	0	0	0	0
15. Football	0.2	0.2	0	0.2	0.2	0
16. Calendar	0.1	0.1	0	0.3	0.1	0.1
17. Software Engineering	0.2	0.3	0.1	0.5	0.3	0.1
18. Student Expo	0.4	0.2	0.2	0.4	0.2	0.2
19. Police	0.2	0.2	0	0.2	0.2	0
20. Xin Ye	0.1	0.1	0	0.1	0.1	0

(3) Discussion

a. The search engine can find relevant pages for 19 out of 20 queries in top 10 results. Only for “Meal Plan” the engine found 0 relevant pages. The reason might be because the crawler did not crawl enough pages that were relevant to “food” or “meal plan”.

b. The larger number of pages that have been crawled, the higher precision of the search engine.

c. The engine in this project performs worse than the Google search engine. Because Google has larger index, and uses more extensions for searching. For example, Google might also combine “title” score and PageRank score for ranking.

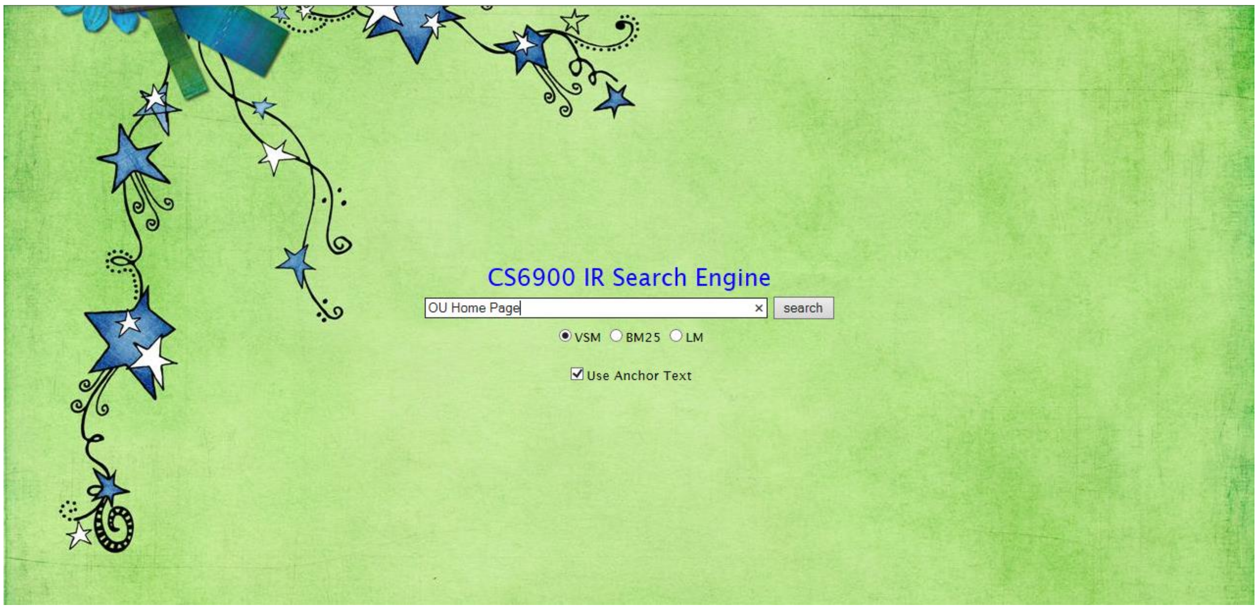
d. Anchor text helps searching. For example, using anchor text helps find relevant pages about “ADA compliance”.

e. VSM performs similarly with BM25. But LM performs worse than VSM and BM25 in this project. The reason may be because the smoothing parameter played a great impact on the LM result. The smoothing parameter needs to be tuned to improve the LM performance.

f. This project also displays snippets of pages, and highlights query terms within snippets. However, for the query “OU Home Page”, the engine does not print any snippets of www.ohio.edu. The reason is because this page does not contain “OU”, “Home Page”. But this page was ranked at the top position because there were many anchor text of “OU Home page” that points to it.

g. Because crawling a huge collection of pages (more than 100K) once needs much time and memory, so it is better to crawl pages step-by-step manually. When using Nutch for crawling, it is better to begin with a small segment first, and crawl an appropriate number of pages each time, and combine together to a large collection at the end.

5. Screenshots of the search engine

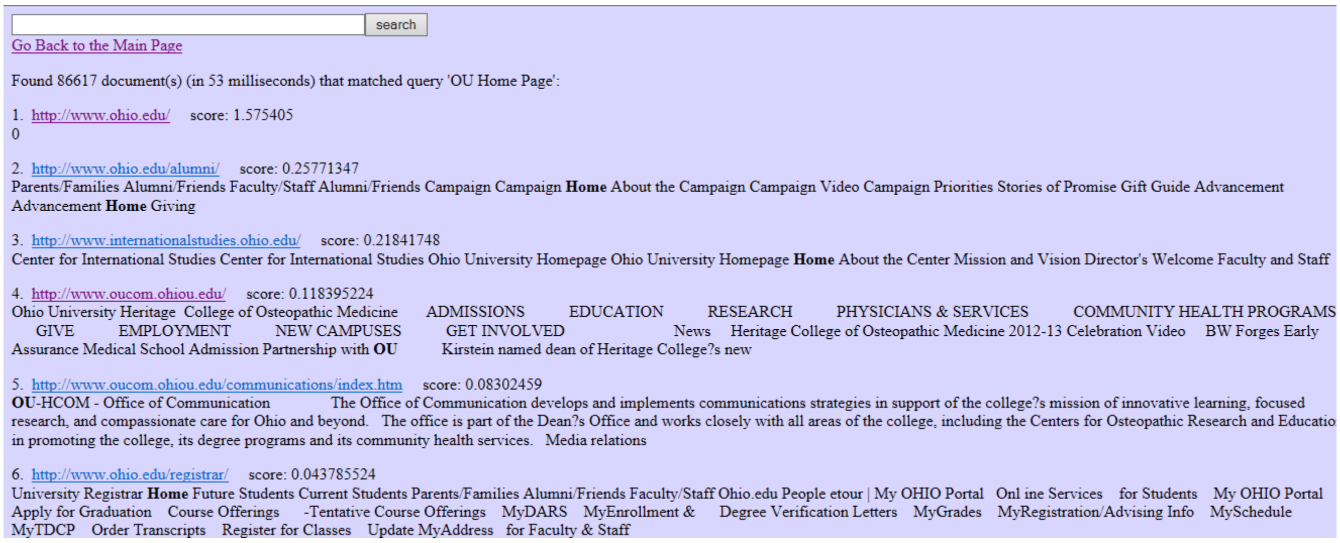

CS6900 IR Search Engine

OU Home Page   search

[✓ VSM] BM25 LM

[✓ Use Anchor Text]

Figure 1 The main page


search

Go Back to the Main Page

Found 86617 document(s) (in 53 milliseconds) that matched query 'OU Home Page':

1. http://www.ohio.edu/ score: 1.575405
0

2. http://www.ohio.edu/alumni/ score: 0.25771347
Parents/Families Alumni/Friends Faculty/Staff Alumni/Friends Campaign Campaign **Home** About the Campaign Campaign Video Campaign Priorities Stories of Promise Gift Guide Advancement Advancement **Home** Giving

3. http://www.internationalstudies.ohio.edu/ score: 0.21841748
Center for International Studies Center for International Studies Ohio University Homepage Ohio University Homepage **Home** About the Center Mission and Vision Director's Welcome Faculty and Staff

4. http://www.oucom.ohiou.edu/ score: 0.118395224
Ohio University Heritage College of Osteopathic Medicine ADMISSIONS EDUCATION RESEARCH PHYSICIANS & SERVICES COMMUNITY HEALTH PROGRAMS GIVE EMPLOYMENT NEW CAMPUSES GET INVOLVED News Heritage College of Osteopathic Medicine 2012-13 Celebration Video BW Forges Early Assurance Medical School Admission Partnership with **OU** Kirstein named dean of Heritage College?s new

5. http://www.oucom.ohiou.edu/communications/index.htm score: 0.08302459
OU-HCOM - Office of Communication The Office of Communication develops and implements communications strategies in support of the college?s mission of innovative learning, focused research, and compassionate care for Ohio and beyond. The office is part of the Dean?s Office and works closely with all areas of the college, including the Centers for Osteopathic Research and Educatio in promoting the college, its degree programs and its community health services. Media relations

6. http://www.ohio.edu/registrar/ score: 0.043785524
University Registrar **Home** Future Students Current Students Parents/Families Alumni/Friends Faculty/Staff Ohio.edu People etour | My OHIO Portal Onl ine Services for Students My OHIO Portal Apply for Graduation Course Offerings -Tentative Course Offerings MyDARS MyEnrollment & Degree Verification Letters MyGrades MyRegistration/Advising Info MySchedule MyTDCP Order Transcripts Register for Classes Update MyAddress for Faculty & Staff

Figure 2 Search result with displayed snippets