IR ASSIGNMENT 3

Q1. STEPS FOLLOWED TO SET UP SOLR AND INDEX WEB PAGES

- Software: Solr 5.3.1 for indexing the crawled web pages
- School crawled: USC School of Cinematic Arts
- Root URL: cinema.usc.edu
- Submitted by Vaishnavi Janardhan, USC ID - 3084414021

1. Installation of Solr and XAMPP

- Solr 5.3.1 was installed according to the instructions provided in the assignment description
- To host the apache server, XAMPP was installed

2. Indexing and Searching

- The crawled data folder (from the previous assignment) was placed under solr/ directory
- A new was created to store the indexed results
    - bin/solr create –c myexample
- Indexing was carried out
    - Index: bin/post –c myexample “path to folder where the files are stored”
- managed-schema file in the conf folder of the the core that we created had to be renamed to schema.xml. The changes were made in schema.xml according to the instructions provided

3. Search engine algorithm comparison

- A pageRankData.csv file was generated: This contains the outgoing links in each of the pages visited during crawling the USC School of Cinematic Arts (from the previous assignment)

- A graph of the incoming and outgoing links was created using the networkx python library

- The initial configuration of the pagerank interface were as follows:
    - alpha = 0.85
    - personalization = None
    - max_iter = 100
    - tol = 1e-06
    - nStart = None
    - weight = 'weight'
    - dangling = None

- The pageRank scores (in the format doc_id=page_rank_score) are stored in the external pagerank.txt file

- Indexed the downloaded files from the previous assignment by using solr

- Modified the schema.xml to work with the newly generated external_pageRankFile.txt. The following were the changes to schema.xml

<fieldType name="external" keyField="id" defVal="0" class="solr.ExternalFileField" valType="pfloat"/>
<field name="pageRankFile" type="external" stored="false" indexed="false"/>

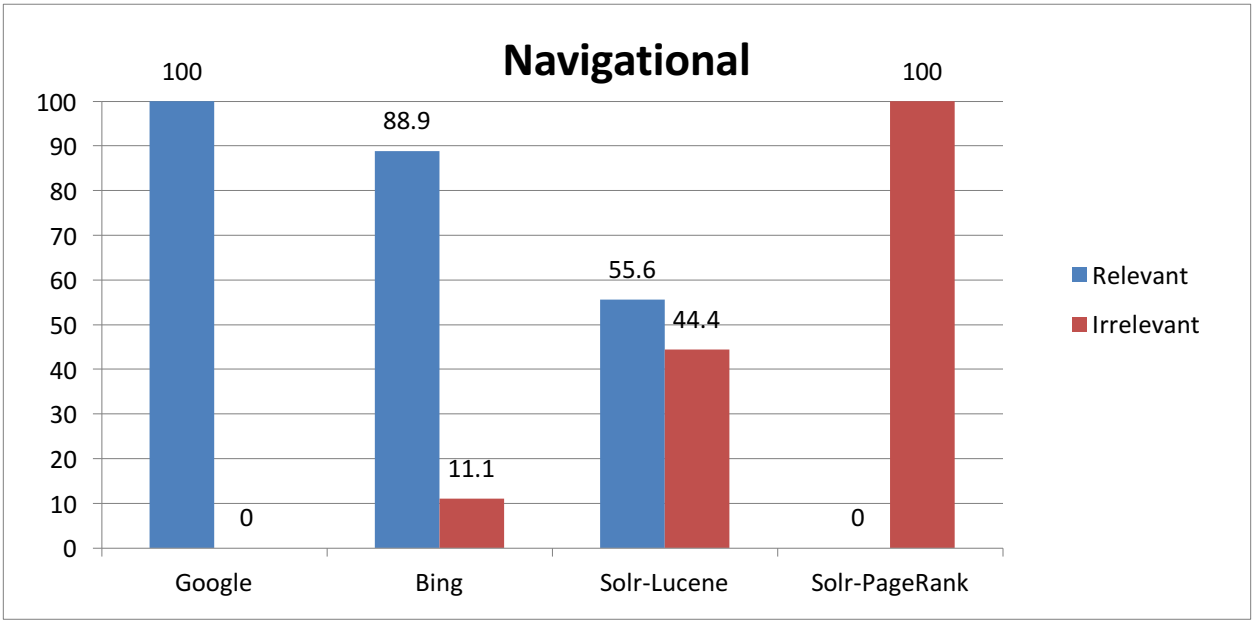
- The following changes were made to solrconfig.xml to support PageRank ordering

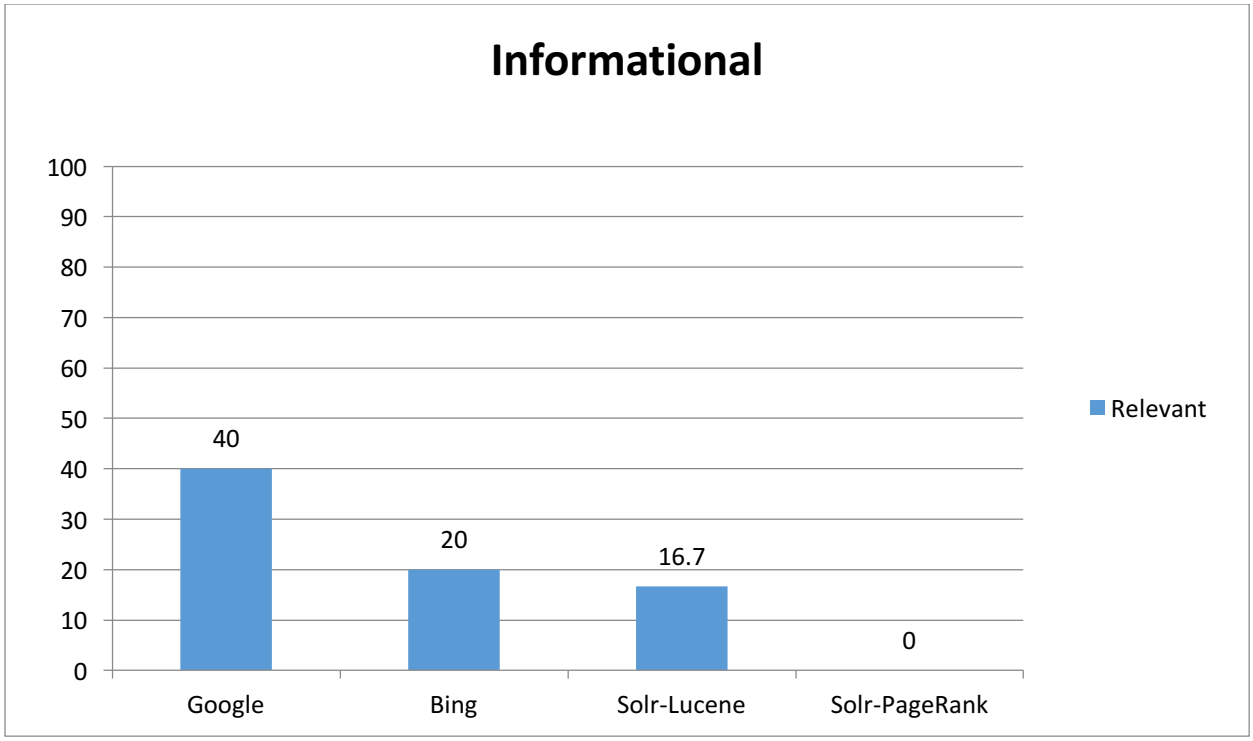
<listenerevent="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader"/>
<listener event="firstSearcher"
class="org.apache.solr.schema.ExternalFileFieldReloader"/>

4. Results

- Reload the core by visiting the Solr Dashboard UI
- A User interface is provided to facilitate the querying. There is an option to choose from solr's internal ranking algorithm or the pagerank algorithm. A PHP program is used to query the results to Solr.
- After the user submits a query, the query is processed and the results are displayed using solr's search API.
- ❖ Queried USC School of Cinematic Arts using the same queries as in the first assignment after starting the Apache web server

Q2. Analysis of the Navigational and Informational queries





- Overall, Google performs the best amongst all the search algorithms
- Comparing Lucene and PageRank, Lucene performs better than PageRank for both Navigational and Informational queries

❖ ***Lucene*** (Default Solr) uses a combination of the Boolean model and the Vector Space model to determine how relevant a given document is to the user's query.

❖ The vector space model is based on term frequency. The Boolean model is used to narrow down on the range of pages that need to be scored based on the use.

❖ When the search engine is queried, the relevance of the documents is obtained by calculating the cosine similarity of the documents and the query.

❖ Solr's Lucene performs quite well, especially for the navigational queries. But it does not outperform Google due to the fact that Solr is indexing a small amount of data as compared to commercial search engines such as Google or Bing that index a large volume of pages.

❖ More relevant results can be obtained by changing the settings to use AND logic on the queries.

❖ ***PageRank*** is one of the algorithms used by Google to determine a relevance of a web page. PageRank uses the information of number of incoming and outgoing links directed to and from a web page to determine the rank of a page.

❖ PageRank ordering on Solr's index is comparatively poor, this can be attributed to the small amount of crawled data and the scope of this particular assignment.

Screenshots for the Query : “Bruce A Block USC School of Cinematic Arts”

1. Lucene (Solr Default) Search Results

Search: Bruce A Block USC School
Search Algorithm to be used:
(•) Lucene(Solr Default) ( ) PageRank Submit

Results 1 - 10 of 2637:

Click to go to the Link
Author:N/A
Date:N/A
Size: 22.45KB
Title: USC Cinematic Arts | School of Cinematic Arts Directory Profile

Click to go to the Link
Author:N/A
Date:N/A
Size: 22.453KB
Title: USC Cinematic Arts | School of Cinematic Arts Directory Profile

Click to go to the Link
Author:N/A
Date:N/A
Size: 22.512KB
Title: USC Cinematic Arts | Directory of SCA Faculty

Click to go to the Link
Author:N/A
Date:N/A
Size: 54.541KB
Title: USC Cinematic Arts | Faculty

Click to go to the Link
Author:N/A
Date:N/A
Size: 56.372KB
Title: USC Cinematic Arts | Directory of SCA Faculty

Click to go to the Link
Author:N/A
Date:N/A
Size: 29.197KB
Title: USC Cinematic Arts | School of Cinematic Arts News

- Lucene returned a few relevant results.

2. PageRank Search Results

Search: Bruce A Block USC School
Search Algorithm to be used:
○ Lucene(Solr Default) ◉ PageRank Submit

Results 1 - 10 of 2637:

Click to go to the Link
Author:N/A
Date:N/A
Size: 40.63KB
Title: USC Cinematic Arts | Home

Click to go to the Link
Author:N/A
Date:N/A
Size: 25.267KB
Title: USC Cinematic Arts | Access

Click to go to the Link
Author:N/A
Date:N/A
Size: 27.444KB
Title: USC Cinematic Arts | Board of Councilors

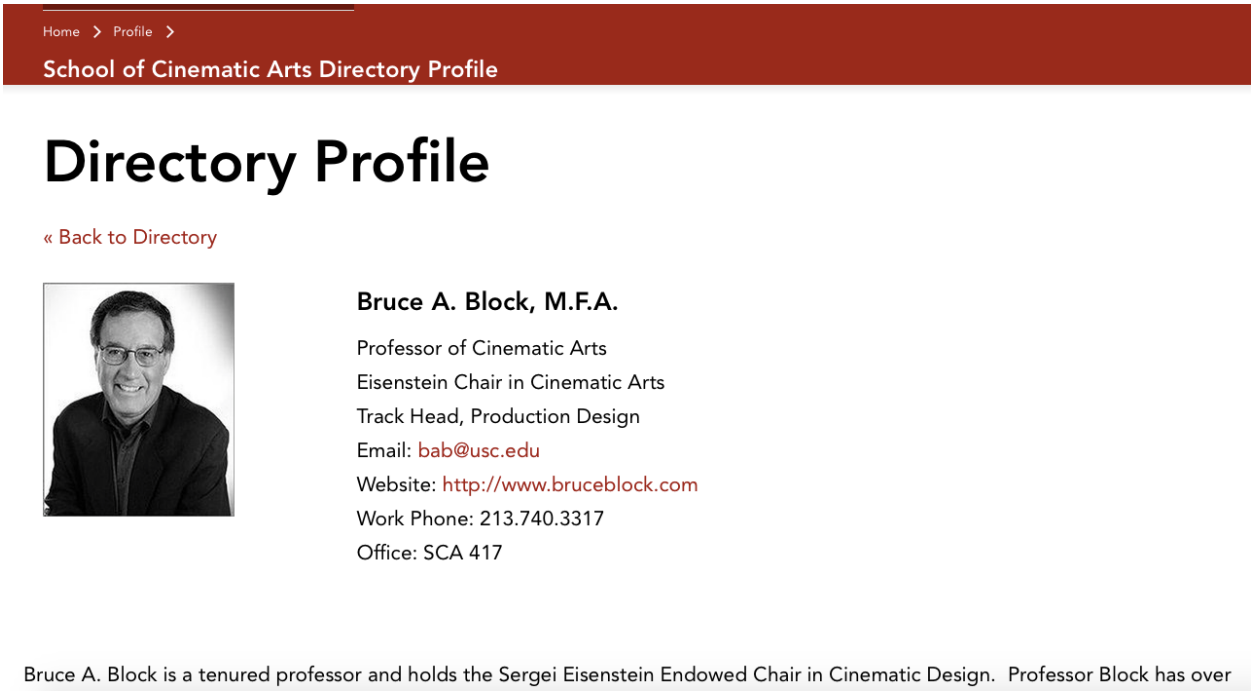
Click to go to the Link
Author:N/A
Date:N/A
Size: 33.469KB
Title: USC Cinematic Arts | Directions & Maps

Click to go to the Link
Author:N/A
Date:N/A
Size: 25.898KB
Title: USC Cinematic Arts | Diversity

Click to go to the Link
Author:N/A
Date:N/A
Size: 29.258KB
Title: USC Cinematic Arts | History from 1929 to 1941

- PageRank returned no relevant results.

3. **After clicking the first link – In Lucene search results**



- The home page of the faculty is obtained