**Homework 3: Comparing Search Engine PageRanking Algorithms**

Name: Vrushali Peshave

USC ID: 4171565740

Steps followed to complete homework:

1. Step 1:Installation

* Installed Ubuntu 12.04 and solr 5.3.1 as per the instructions given in pdf.

1. Step 2:Creation of core in solr and indexed all downloaded files

Commands Used:

* Bin/solr start:To start solr(By default runs on 8983 port)
* Bin/sor create –c mycore3:To create a new core named “mycore3”
* Uncommented the following content in **managed-schema** file

<copyField source="title" dest="\_text\_"/>

<copyField source="author" dest="\_text\_"/>

<copyField source="description" dest="\_text\_"/>

<copyField source="keywords" dest="\_text\_"/>

<copyField source="content" dest="\_text\_"/>

<copyField source="content\_type" dest="\_text\_"/>

<copyField source="resourcename" dest="\_text\_"/>

<copyField source="url" dest="\_text\_"/>

* **bin/post -c mycore3 /home/vrushali/Desktop/InformationRetrieval/CrawledData/**

**Downloads**: To index all downloaded files in Solr.

All the files which are downloaded as part of HW2 are put up in the above mentioned system path. The file name of each downloaded files is kept as the entire URL after replacing /l symbol with ‘\_’ symbol and removing http: or https: and remaining special characters and extension given as .html

For example,

<http://gould.usc.edu/about/aba-required-disclosures/gpa-and-lsat-scores/> becomes **gould.usc.edu\_about\_abarequireddisclosures\_gpaandlsatscores\_.html**

* Uncomment the following content in solconfig.xml file under the **“select” requestHander** section **<str name="df">\_text\_</str>**

1. Step 3:Calculation of pagerank and creation of external\_pageRankFile:

* As a part of HW2 assignment, all the downloaded URLS are stored in pagerankdata.csv along with all of its outgoing URLS.(1st column is url of the downloaded html file and subsequent columns contain outgoing links) Then I wrote a python script to calculate pageRank based on this csv file using networksx library.
* I created Digraph graph G in which all urls in pageRankFile are nodes and edge is formed between a url and its every outgoing link.
* After creating the graph G, It was passed as an argument to networkx.pagerank() function
* Value of alpha passed is 0.85 as it is proven experimentally that this value gives best performance.
* Total nodes in a graph are 14483 Hence I set max\_itr as log(14483) =4(approx) and pagerank algorithm will converge after 4 iterations.
* The pageRank function returns the calculated pageRank for each URL as a dictionary with key as a url and value of pagerank as a value

<http://gould.usc.edu/about/aba-required-disclosures/gpa-and-lsat-scores/=0.000055>

Even though the page rank is calculated for each URL, for Solr to understand it, we should convert the URLS to Solr Document IDS and write it into external\_PageRankFile.txt. Solr Document ID is an address of downloaded url which is indexed.So along with pageRankFile ,I have created one more file which stores key as a url and value as its syatem path

<http://gould.usc.edu/about/aba-required-disclosures/gpa-and-lsat-scores/>-> **/home/vrushali/Desktop/InformationRetrieval/CrawledData/Downloads/gould.usc.edu\_about\_abarequireddisclosures\_gpaandlsatscores\_.html**

This mapping helps to go to actual link when search results are obtained from solr.

When pageRank is obtained (output is of the form of dictionary) it is written to external\_pagerankFile.txt as **/home/vrushali/Desktop/InformationRetrieval/CrawledData/Downloads/gould.usc.edu\_about\_abarequireddisclosures\_gpaandlsatscores\_.html=0.00005**

1. Step 4:Configuration in solr to use external\_pageRankFile

Once the **external\_PageRankFile.txt** is created, It is been added into data folder of the core ‘mycore3’

And then the following lines were added in manage-schema.xml

<fieldType name="external" class="solr.ExternalFileField" keyField="id" valType="pfloat" defVal="0"/>

<field name="pageRankFile" type="external" indexed="false" stored="false"/>

The following two lines were added in solconfig.xml

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

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

Than,I reloaded the core in core\_admin section in order to reflect the changes.

1. Step 5:Design front end to search for a query

I used given PHP code to design User Interface then added two radio buttons as Search using Solr and Search using pagerank.

When search using solr is selected default pageranking algorithm of solr is used to fetch the results while when search using pagerank is selected external pageranking file is used to calculate pagerank of the pages which is achieved by passing additional prameters to search function.

When Solr returns the top results, I display the following details in a table

1) Title 🡪 $doc->title

2) Author 🡪 $doc->author

3) Date Created 🡪 $doc->created

4) Size 🡪 ($doc->stream\_size)/1024 in KB

5) and a link to the original page. 🡪 $doc->id

$doc->id will only give Solr id instead of actual URL, I used pagerankmap.csv to obtain mapping of URL for the given Solr id.

1. Step 6:Query Results and Analysis

I used the same queries used in Assignment#1 to analyze search results.Results obtained are as follows:

Navigational:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Google | Bing | Solr | PageRank |
| Relevant | 1 | 0.75 | 0.25 | 0 |
| Irrelevant | 0 | 0.25 | 0.75 | 1 |

Informational:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Google | Bing | Solr | PageRank |
| Relevant | 0.266667 | 0.2 | 0.266667 | 0.133334 |
| Irrelevant | 0.733333 | 0.8 | 0.733333 | 0.866666 |

**Comparison of various page ranking algorithms:**

**Analysis of relevancy results:**

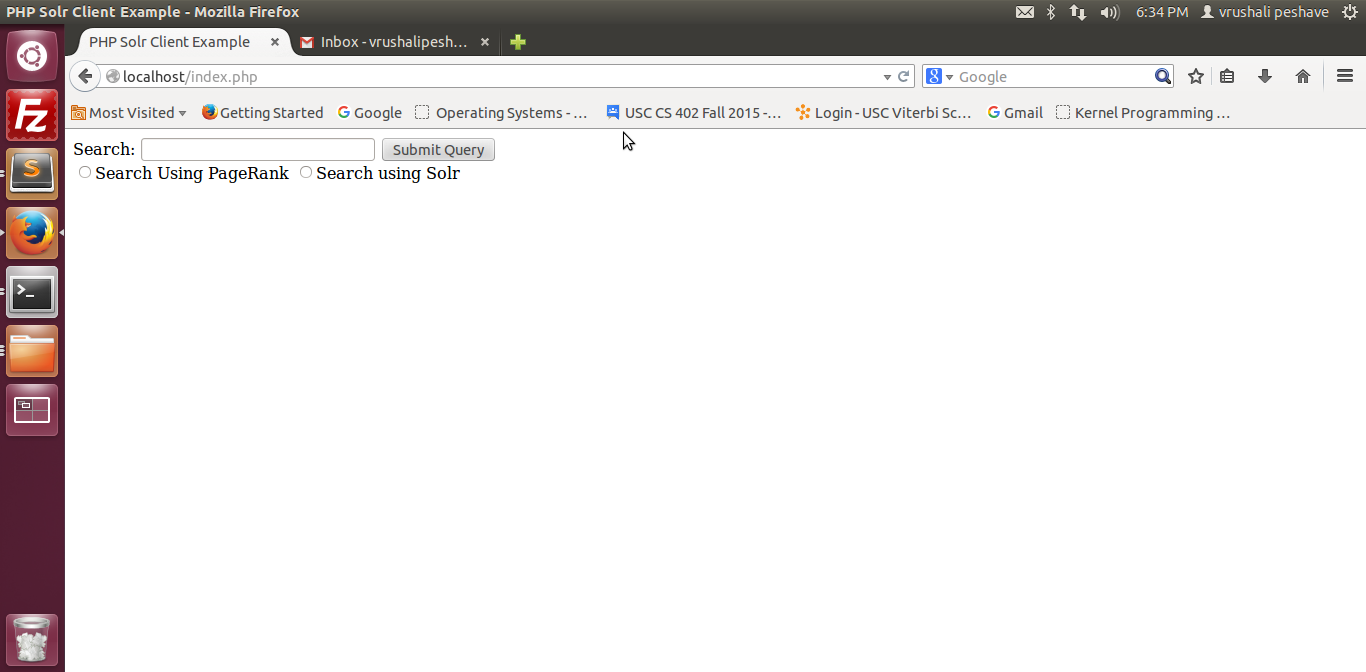
* In Informational and Navigational queries google performs best as compared to all other. PageRank is only **one of numerous methods Google uses** to determine a page’s relevance or importance. Along with pagerank it takes into consideration the quality of the page that is how relevant the document is w.r.t query.
* Bing performs well for the above same reasons.
* For Informational queries performance of Google, Bing and solr are comparable.
* Solr uses lucene tf.idf (term frequency-inverse document frequency) scoring model adds some more features to it. It performs better than external pagerank algorithm as external pagerank doesn’t give weightage to content relevancy.

**Why some pages have higher page Rank:**

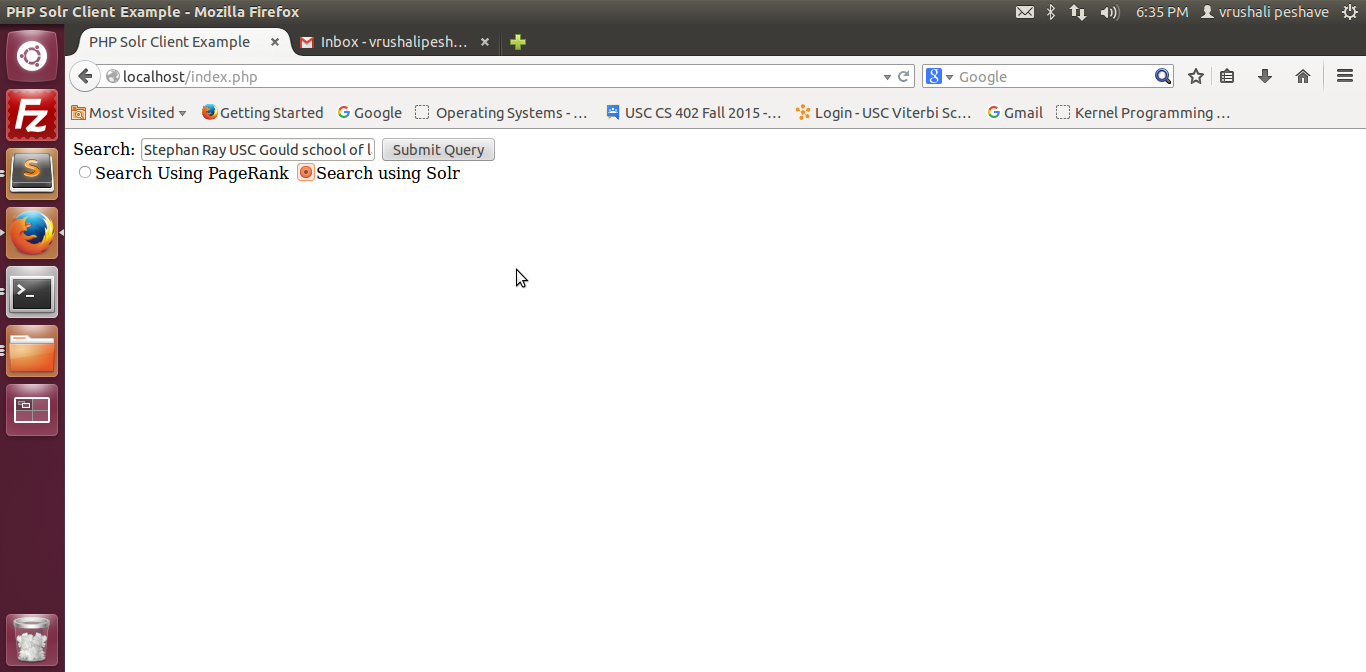
* As solr uses lucene scoring tf.idf model, it can be seen that more the term frequency of query term in a document more is the score of the document. So documents with high query term frequency have high page rank. Order of query terms is not taken into consideration thus solr gives poor performance in case of navigational search queries(Most relevant document is not obtained at top position )
* Also solr takes into consideration relevance of content w.r.t query terms documents it give more precision compared to external pageRank which considers only incoming and outgoing links of the page.
* It is observed that solr gives more relevant documents thus number of documents retrived for a particular query is less that number of documents retrieved by external pagerank algorithm.
* External pagerank does not take document-query similarity into consideration so irrelevant documents are obtained. PageRank of page is only decided by incoming and outgoing links of the page.

Screenshots of the execution of query result:

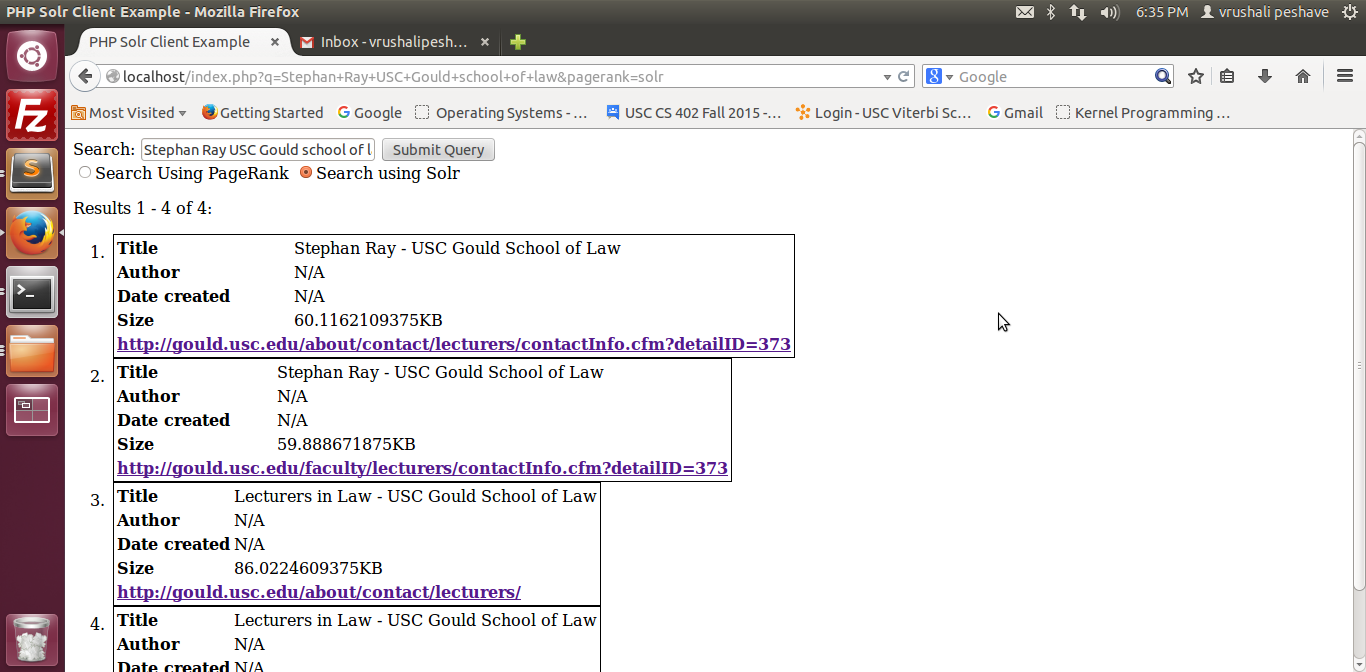
1. User Interface:



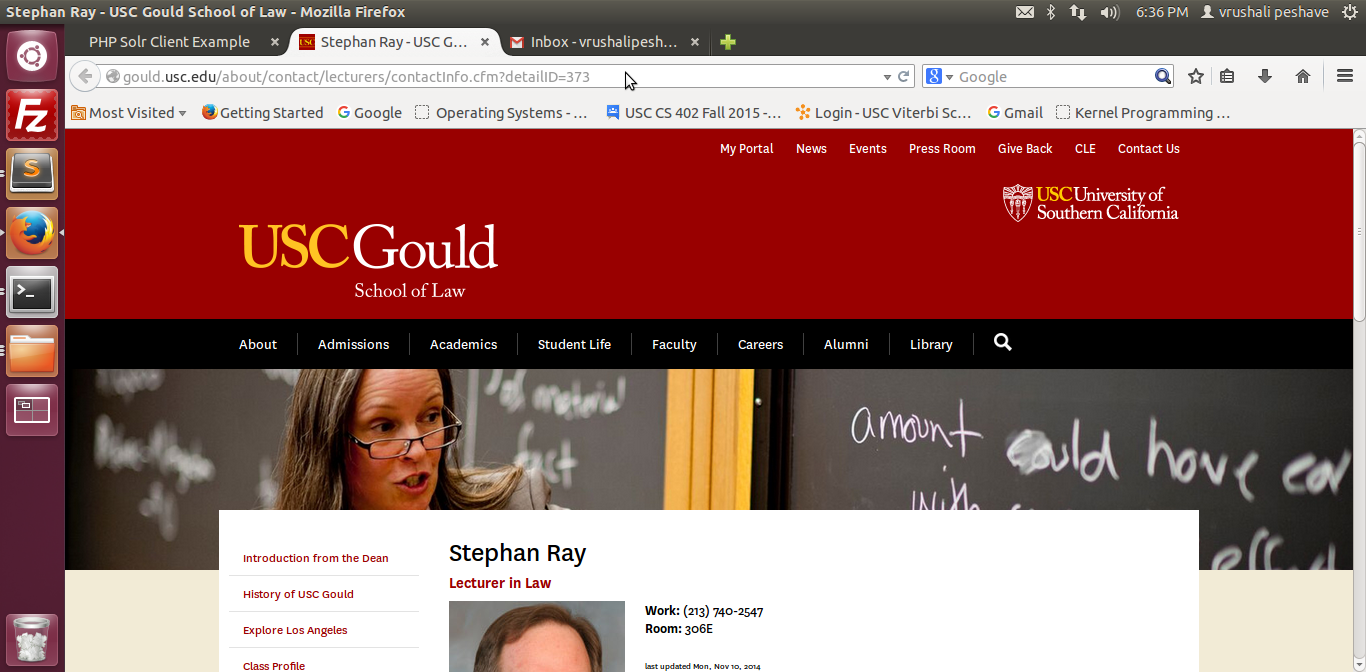
2. Before executing query “Stephan Ray USC Gould school of law” in solr



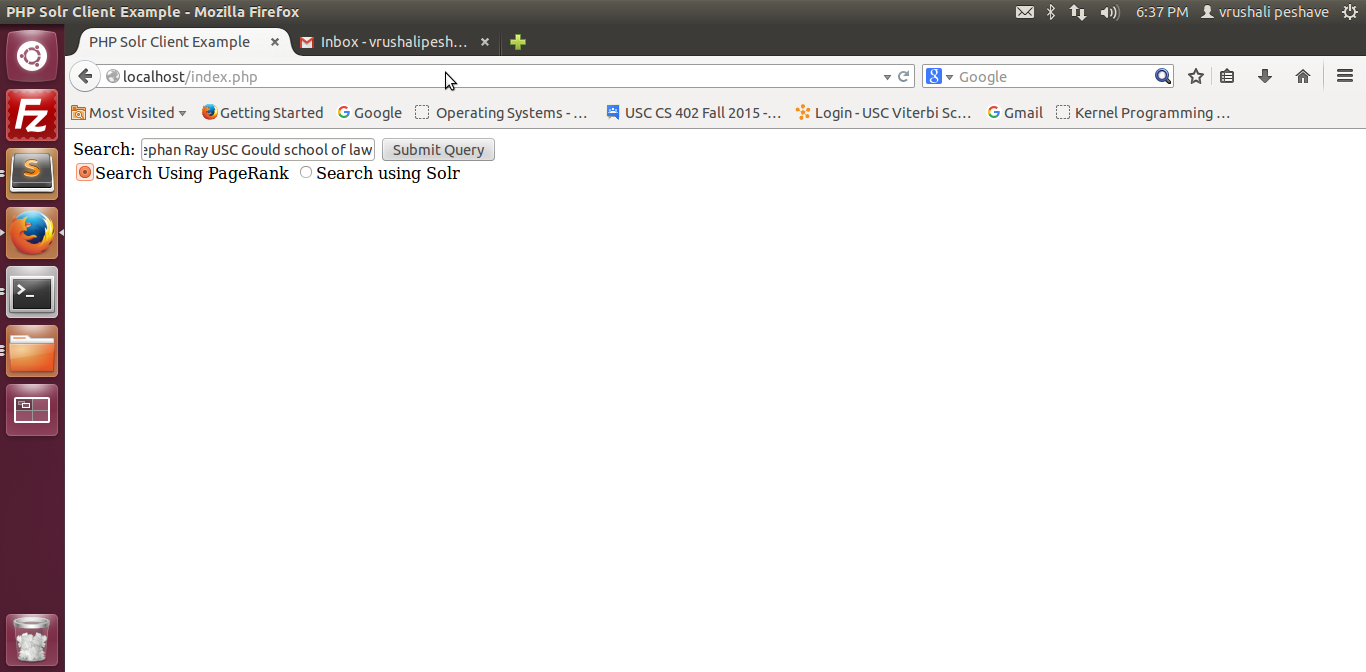
3.Search results of query “Stephan Ray USC Gould school of law”



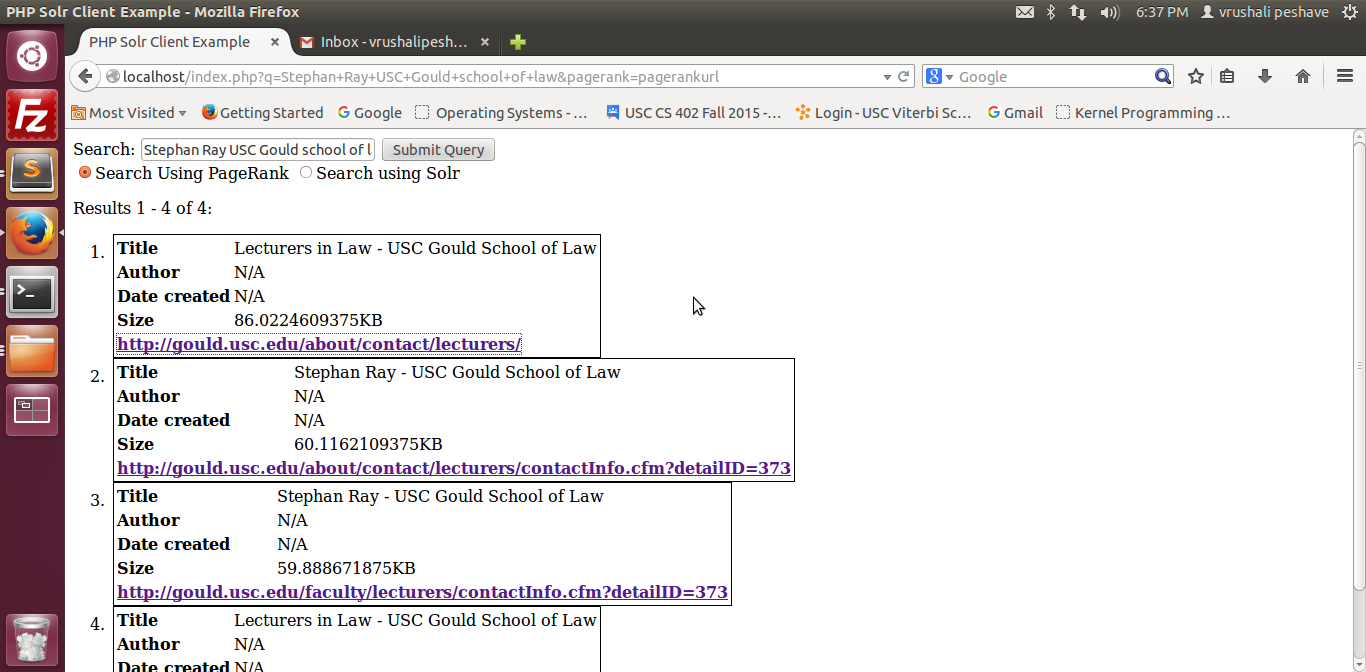
4.Redirected to 1 st link of the search results:



5. Before executing query “Stephan Ray USC Gould school of law” in pagerank



6. Search results of query “Stephan Ray USC Gould school of law”



7. Redirected to 1 st link of the search results

