<u>SEARCH ENGINE WITH PAGE RANK</u>

By,

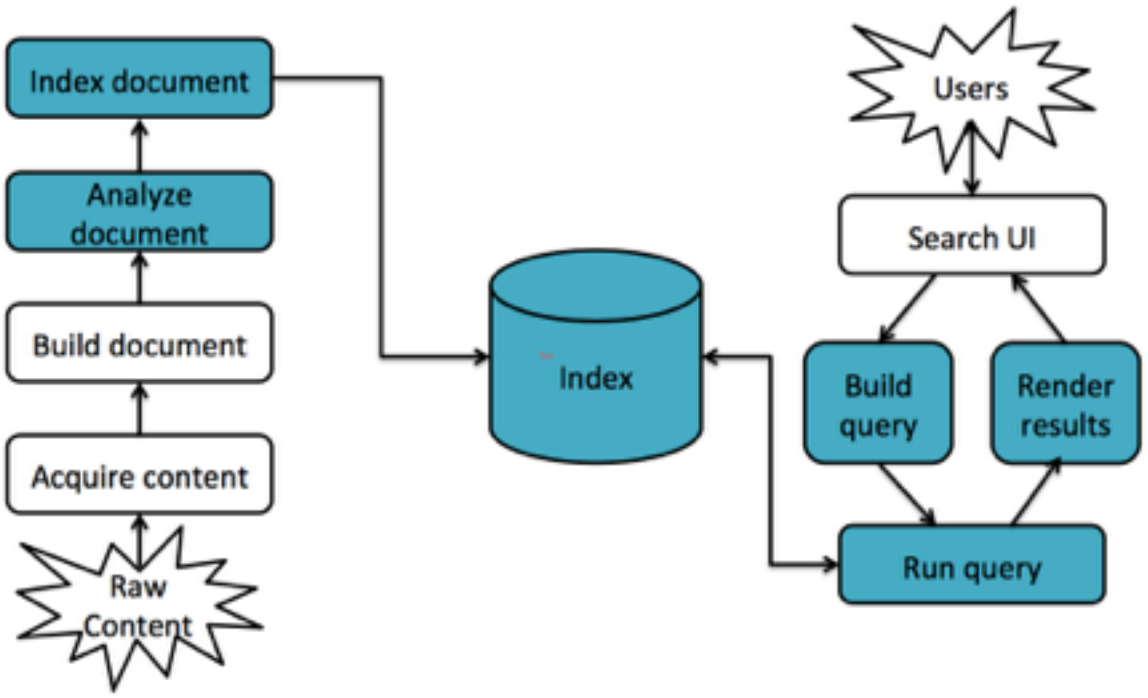
Piyush Mangrulkar - 861245524
Vikash Kumar - 861245614

INTRODUCTION

The amount of information on the web is growing rapidly. Search engines index tens to hundreds of millions of pages involving a comparable number of distinct pages.The goal of a search engine is to optimize the speed of the query, and find the top documents where the query term occurs. Without a pre-computed index, the search engine would have to scan every document in the corpus, which would require considerable time and computing resources.

OVERVIEW

The **Index creation** process is an important phase in the lifecycle of a search engine. The indexer collects, parses, and stores data to facilitate fast and accurate information retrieval. The purpose of storing an index is to optimize speed and performance in finding **relevant** documents for a search query. Many search engines tend to incorporate an **inverted index** when evaluating a search query to quickly locate documents containing the words in a query and then rank these documents by relevance.



ARCHITECTURE

The Input to Phase-II of the Search Engine is the crawled data, from Phase-I. All the pages are parsed and indexed for ease of searching later on. Boost is applied to important fields, such as title while indexing the documents. Page Rank is calculated for all the pages and saved to disk. Based on the search query user has typed, the index will be searched and Top documents according to score will be retrieved, along with the snippet for that page. Also, page rank can be incorporated in document scoring, depending on user choice. Clicking on any of the results will direct the user to that page.

THE CRAWLING / DATA COLLECTION STRATEGY

-Indexing Pages

i) Navigate to the Directory containing downloaded files
ii) Parse each file in the directory and extract all the required information like title, body, metadata , etc
iii) Set the boost to required fields to increase their importance
iv) Add the info(both indexable and not) to a Lucene document, and add this document to the index

-Searching For Pages

i) Accept search query from the user
ii) Parse the tokens from the query and assign weights to fields(title, metadata) for retrieving results
iii) Retrieve the top docs by searching in the index
iv) Retrieve the URL for the top documents
v) Generate the snippet for the top docs
vi) Navigate to the website which user clicks

-Page Rank

i) Parse all the documents and get all the outgoing hyper link associated with it.
ii) Make a Graph representing all the pages as vertex of the graph and all the hyper link as edges of the graph.
iii) Representing the above Graph in Adjacency list form.
iv) Make one more Adjacency representation of Graph for the page which is linked by other pages. i.e all the pages which have at least one hyper link to the particular node as these pages will contribute in the page rank calculation of the particular node/vertex
v) Page Rank equation
PageRank(A) = (1 - d) N + d * (PR(B)/no of outgoing link + PR(C)/no of outgoing link + ….. + PR(n)/no of outgoing link)

where d = damping factor
N = total number of pages in the system.

All the incoming nodes on the vertex can be found by IncomingLink Adjacency list. and number of out going link for that page can be found by main graph Adjacency list

vi) Convergence factor
Page rank will be calculated / iterated when the successive difference between pages will be grater than convergence factor

-Snippet Generation

i) Check for the position of query term in the result document
ii) Define CharsCount as no. of character to display in the snippet
iii) If term appears in the beginning of the document, display snippet from document starting to CharsCount
iv) If the term appears in the end of the document, display snippet from (document.length - CharsCount) to end of document
v) If the term appears in the middle of the document, display snippet as CharsCount no of characters before and after the position at which query term occurs
vi) If the Document length is less than CharsCount, then display the document as the snippet

COLLABORATION

i) Index Creation - Piyush
ii) Page rank implementation - Vikash
iii) Search - Piyush
iv) Web Server Implementation - Vikash
v) Snippet generation - Piyush
vi) User Interface - Vikash
vii) Unit Testing - Vikash & Piyush
viii) Report - Piyush

LIMITATIONS

i) Page Rank will be 1 by default for the pages which have not been crawled, due to 5GB memory limit

ii) Lucene Limitations
a) Maximum number of unique terms in any index segment is ~2.1 Billion, as Lucene’s current implementation uses a JAVA int to hold the term index.

b) Similarly Lucene uses a Java int to refer to document numbers and index file format uses an Int-32 to refer to document numbers, so it also has an upper limit.

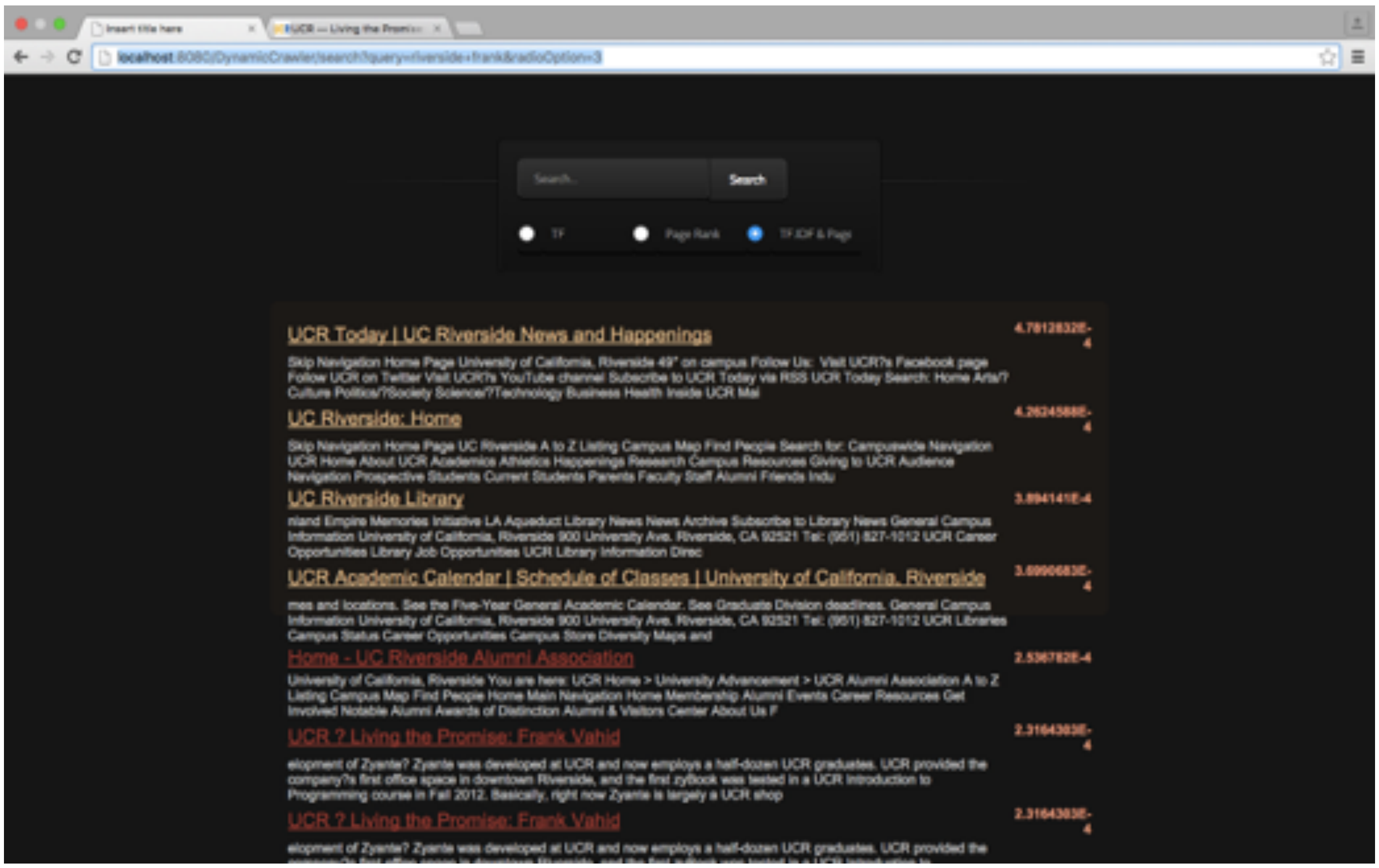
EXTRA FOR BONUS POINTS

i) Snippets are computed for each web pages in the results, depending on where the search term occurs in the page
ii) Page Rank is implemented, and user is provided with option to use it
iii) Certain fields, such as title have been boosted to indicate weightage

DEPLOYMENT INSTRUCTIONS

i) Start the Tomcat Server
ii) Double click in .sh file
iii) Enter Index directory and downloaded path of crawled data (downloaded HTMLs)

SCREENSHOT



localhost:8080/DynamicCrawler/search?query=riverside+frank&radioOption=3

TF [✓ TF IDF & Page] Page Rank

UCR Today | UC Riverside News and Happenings 4.78128332E-4
Skip Navigation Home Page University of California, Riverside 49° on campus Follow Us: Visit UCR?s Facebook page Follow UCR on Twitter Visit UCR?s YouTube channel Subscribe to UCR Today via RSS UCR Today Search: Home Arts/?Culture Politics/?Society Science/?Technology Business Health Inside UCR Mai

UC Riverside: Home 4.26245888E-4
Skip Navigation Home Page UC Riverside A to Z Listing Campus Map Find People Search for: Campuswide Navigation UCR Home About UCR Academics Athletics Happenings Research Campus Resources Giving to UCR Audience Navigation Prospective Students Current Students Parents Faculty Staff Alumni Friends Indu

UC Riverside Library 3.8941416E-4
nland Empire Memories Initiative LA Aqueduct Library News News Archive Subscribe to Library News General Campus Information University of California, Riverside 900 University Ave. Riverside, CA 92521 Tel: (951) 827-1012 UCR Career Opportunities Library Job Opportunities UCR Library Information Direc

UCR Academic Calendar | Schedule of Classes | University of California, Riverside 3.69906683E-4
mes and locations. See the Five-Year General Academic Calendar. See Graduate Division deadlines. General Campus Information University of California, Riverside 900 University Ave. Riverside, CA 92521 Tel: (951) 827-1012 UCR Libraries Campus Status Career Opportunities Campus Store Diversity Maps and

Home - UC Riverside Alumni Association 2.5367828E-4
University of California, Riverside You are here: UCR Home > University Advancement > UCR Alumni Association A to Z Listing Campus Map Find People Home Main Navigation Home Membership Alumni Events Career Resources Get Involved Notable Alumni Awards of Distinction Alumni & Visitors Center About Us F

UCR ? Living the Promise: Frank Vahid 2.31643383E-4
elopment of Zyante? Zyante was developed at UCR and now employs a half-dozen UCR graduates. UCR provided the company?s first office space in downtown Riverside, and the first zyBook was tested in a UCR introduction to Programming course in Fall 2012. Basically, right now Zyante is largely a UCR shop

UCR ? Living the Promise: Frank Vahid 2.31643383E-4
elopment of Zyante? Zyante was developed at UCR and now employs a half-dozen UCR graduates. UCR provided the