



SEARCH ENGINE EVALUATION

COMPSCI 744: Text Retrieval and its Applications in Biomedicine

Instructor

Dr. Susan Mcroy

Jay Tank

jaytank@uwm.edu



Search Engine Evaluation

Table of Contents

1. Introduction
2. Review of Literature
3. Existing System
4. Project Plan
5. Project Resources
6. Project Design
7. Project Implementation
8. Conclusion and Future Scope
9. References

1. INTRODUCTION

With the magnificent amount of information present on web, it is very important to identify whether the search engine satisfy all the requirements of users by their search results. So, it is necessary to evaluate search engines based on user point of view. Basically, evaluation of search engines is a process of determining how well the search engines meet the information need of users. As part of this project, I have created a search engine evaluation system which is based upon page level keywords.

Page level keywords are the keywords found in individual pages of a website. Page level keyword is an important factor to measure the relevancy of the search engine results. The result set retrieved by search engines are containing a huge number of useless web pages.

Users may have to sift through dirt's in order to find gemstones or to rethink his query. So, my work can be a basis to provide more relevant search results to the users. Three Search engines Google, Yahoo and Bing are evaluated based on biomedical and health queries in accordance with page level keywords and other factors.

The system does not tell which search engine in particular gives the best result as it depends on the user's search query and information need but it will give the best results after evaluating the results retrieved from all the search engines that closely meets the information need of the user.

2. REVIEW OF LITERATURE

Research Paper

'RatioRank: Enhancing the Impact of Inlinks and Outlinks' - In this paper a new page ranking algorithm known as the RatioRank is discussed, in which the inlink weights and outlink weights are used with the consideration of number of visit count and is compared with some algorithms by using certain parameters. This algorithm is used to rank the pages retrieved by the search engine.

I have read many other research papers the names of which have been given in the References page. These papers have helped me understand the existing systems and also the concepts that I thought of implementing to make a better system to give more accurate results. These papers have also helped me understand different ways of ranking a page.

PageRank

PageRank is a link analysis algorithm used by Google Search and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references.

A PageRank results from a mathematical algorithm based on the web graph, created by all World Wide Web pages as nodes and hyperlinks as edges, taking into consideration authority hubs such as cnn.com or usa.gov. The rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support. The PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it ("incoming links"). A page that is linked to by many pages with high PageRank receives a high rank itself.

Currently, PageRank is not the only algorithm used by Google to order search results, but it is the first algorithm that was used by the company, and it's best known

3. EXISTING SYSTEM

Most of the current existing systems counts the number of times keywords that are present in the retrieved pages and accordingly comes up with a decision as to which search engine provides the most relevant results. The pages are being ranked just based on the keyword counts and other parameter for ranking the page are not taken into consideration and used needs to fire the same query on different search engine, compare the result obtained from the different search engine and then trust the information.

However, there are some intelligent, advanced and highly complex systems that considers multiple factors for ranking the page and gives user the most accurate results. In my project I have tried to demonstrate the working of such systems on a small scale.

PITFALLS IN EXISTING SYSTEM

The current system for evaluation of search engines suffers from the following pitfalls:

Incorrect Results

Since the system is only using keywords as the criteria to judge search engines it could many times lead to inappropriate results. It may happen that a given page consists a keyword more number of times but is not relevant to the user or is not a reputed page. In this case the current system will give incorrect results.

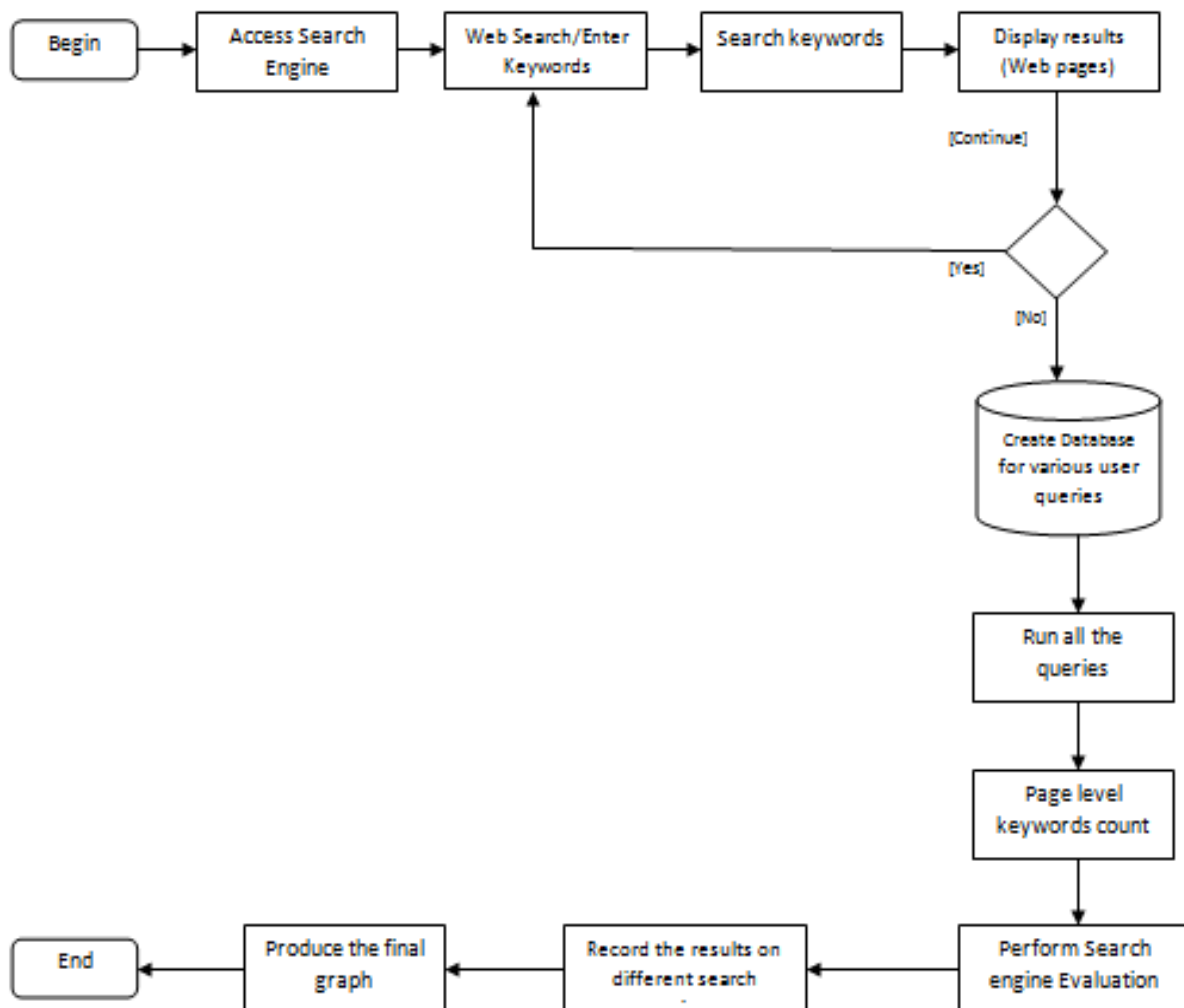
Human Errors

There may be chances that the user enters incorrect keywords during entering the keywords and since the system is completely based on keywords it will provide incorrect results.

4. PROJECT PLAN

The primary idea of this project is to perform evaluation of various popular search engines on the basis of a variety of factors or parameters. The user first enters the search query in the search bar provided. This search query is performed on all search engines under consideration. The pages retrieved are then evaluated on basis of various parameters. They are first parsed for page level keywords. Then they are scored on the basis of inbound and outbound links. There are also other factors like Alexa rank, domain age, Dmoz listing etc which are considered. The final score of each page is calculated by considering all factors. The final results are then displayed

Flowchart



5. PROJECT RESOURCES

Hardware Requirements

1. Processing requirements:
 - P4 and Higher
2. Memory
 - 1 GB RAM
 - 2 GB Hard Disk

Software Requirements

The following are the software development tools

- XAMPP server or any PHP container
- JavaScript enabled Web Browser
- MySQL

Programming languages

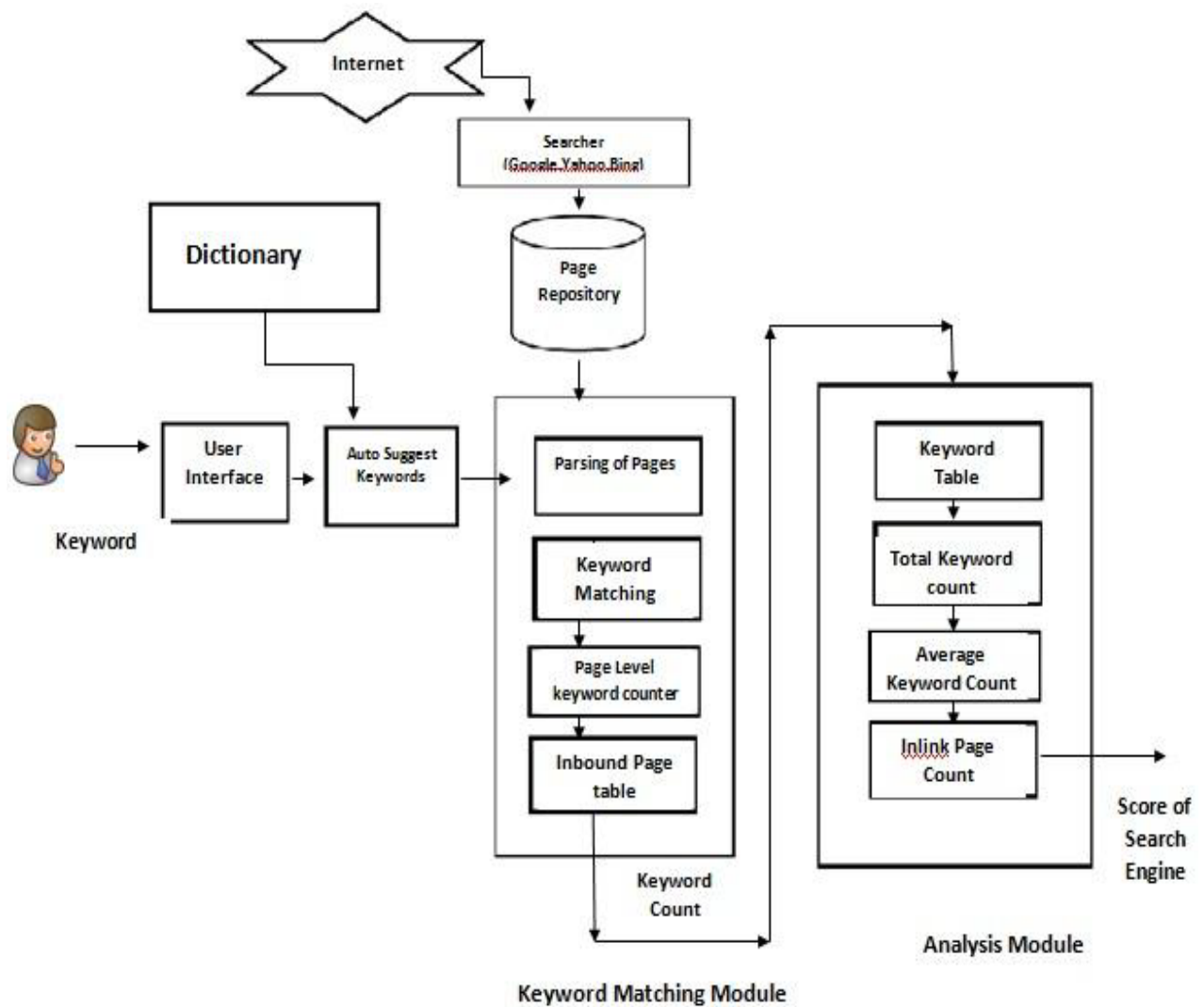
- PHP
- MySQL

Operating Environment

- Windows 10
- Linux
- Mac OS

6. PROJECT DESIGN

System Architecture



The system intends to take search queries as text input from the user. The user will enter the query in the text box provided. The system will then search these queries on different search engines such as Google, Yahoo and Bing and retrieve the pages given as search results from all three search engines. These pages will then be analyzed based on several different factors according to which they will be ranked and hence the search engines will be evaluated on the basis of the pages retrieved. The retrieved pages will first be parsed to check for page level keywords. Page level keywords are the keywords found in individual pages of a website. Page level keyword is an important factor to measure the relevancy of the search engine results. The result set retrieved by search engines are containing a huge number of useless web pages. Users may have to sift through dirt's in order to find gemstones or to rethink his query. So, my work can be a basis to provide more relevant search results to the users. After being evaluated for page level keywords the pages will be evaluated on the basis of inbound and outbound links. There are two types of links: inbound and outbound. Outbound links start from your site and lead to an external site, while inbound links or backlinks, come from an external site to yours. e.g. if cnn.com links to yourdomain.com, the link from cnn.com is a backlink (inbound) for yourdomain.com, however the link is an outbound link from cnn.com's perspective. Backlinks are among the main building blocks to good Search Engine Optimization (SEO). The number of backlinks is an indication of the popularity or importance of that website. Backlinks are important for SEO because some search engines like Google, give more credit to websites that have a large number of quality backlinks, and consider those websites more relevant than others in their results pages for a search query. Hence the pages will be evaluated on the basis of inbound and outbound links for more accurate results. Moreover, I am also using other third-party ranking systems like Alexa rank, Dmoz listing etc. and other parameters as domain age etc. as parameters to rank pages in my system. After considering all the parameters for all pages a final score is generated for each page and the pages are ranked according to this final score.

7. PROJECT IMPLEMENTATION

Module Description

- **Keyword Categorization:** This module takes the keyword from the input Query. Keyword counts are stored in a local database. "Score.php" includes and implements this functionality by means of "count()" method.
- **Archive _trustworthy_results:** These modules summarize the trustworthy result and store it in the database in the archive format. It is a part of "save-result.php" where it is implemented by means of PHP/MYSQL.
- **View_Result:** This module will display the trust worthy score based on the result received from the archived database for the keywords categorized in the first phase. This schema will be used for analysis in the first filtration of keyword.
- **Keyword and Relevance Detection:** Keywords are transmitted to Third party API for natural language processing which acts as the second filter to display the final Score.

Module wise algorithm

- **Keyword Categorization:**
 1. Stream of the all the data retrieved from web search is categorized based on keywords.
 2. The stream is categorized based on 50 health related keywords and then results are displayed on php form.
- **Archive _trustworthy_results:**
 1. A new database schema is created using the old database to store the key parameters for displaying. Hence, it stores all results in a database schema which may be used for further processing
 2. It is a part of "save_result.php" which saves all the retrieved stream data.
- **View_Result:**
 1. View_Result will display contents of "archive.php".It will basically display all saved results
 2. The stream is categorized based on 50 health related keywords and then results are displayed on php form.

- **Keyword and Relevance Detection:**


1. Whenever user fires a search query it will extract all information from GOOGLE, BING, YAHOO.
2. Based on the Factor rating, it will display only those web pages with higher factor.
3. On basis of the relevance factors, webpages are filtered for the final output.

Working with Screenshots



Home Page

Welcome in Google - Truth Discovery On Web Admin



Enter search query what you like to Search from Google

Truth Discovery On Web - Administrative Control Panel Logged in User : admin | [logout](#)

General Settings

- Home Page
- Manage Admin
- Add User
- Logout

Site Admin

- Manage Factor
- Trustworthy Checker
- Search Archive

DB Setting

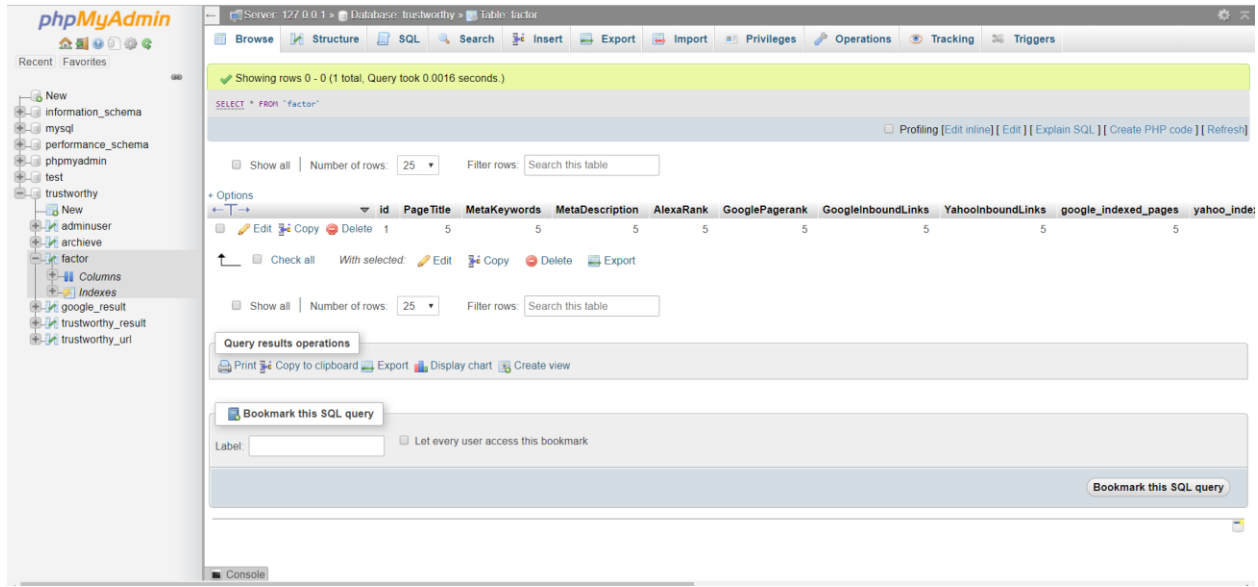
- Truncate Table

FACTOR ALLOCATED SCORE

FACTOR ALLOCATED SCORE

Available Percent: % Max Percent: %

Page Title :	<input type="text" value="5"/>
Meta Keywords:	<input type="text" value="5"/>
Meta Description:	<input type="text" value="5"/>
Alexa Rank:	<input type="text" value="5"/>
Google Pagerank:	<input type="text" value="5"/>
Google Inbound Links:	<input type="text" value="5"/>
Yahoo Inbound Links:	<input type="text" value="5"/>
Google Indexed Pages:	<input type="text" value="5"/>
Yahoo Indexed Pages:	<input type="text" value="5"/>
Domain Age:	<input type="text" value="5"/>
Wot Rating:	<input type="text" value="10"/>
Last Modified Date Of Page:	<input type="text" value="5"/>
Alexa Inbound Links:	<input type="text" value="5"/>
Dmoz Listing:	<input type="text" value="5"/>
Site Advisor Rating:	<input type="text" value="10"/>
Bing Indexed Pages:	<input type="text" value="5"/>
Bing Inbound Links:	<input type="text" value="5"/>
Ask Indexed Pages:	<input type="text" value="5"/>



8. CONCLUSION AND FUTURE SCOPE

The following are the key features of this application:

1. Provides top link from multiple search engines.
2. Stores trustworthy score locally and deletes them after processing to maintain privacy.
3. Easy and convenient to use.
4. Supports cross platform operations (windows, Linux, Mac)

Future Scope

1. Non-English or mixed English language support.
2. Customized results for each user based on the log history and cookies.
3. Future Scope
4. Further development of the Mobile Application can be done for smart phone and other portable devices.
5. Sharing the search results with other users from anywhere around the world.

9. REFERENCES

1. S.Brin and L.Page, “The Antonomy of a Large Scale Hypertextual Web Search Engine,”7th Int.WWW Conf. Proceedings,Australia ,April 1998.
 2. M.Marchiori, “The Quest for Correct Information on the Web: Hyper Search Engine,” The Sixth International Conference(WWW97). Santa Clara, USA,1997.
 3. J.Kleinberg,”Authoritative Source in a Hyperlinked Environment,”Proc.ACM-SIAM Symposium on Discrete Algorithm,1998, pp. 668-677.
 4. K.Bharat, A. Broder, M.R. Heizinger, P.Kumar and S.Venkatasubramanian, “ Connectivity Server: Fast Access to the Linkage Information on the Web,” Proc. 7th International WWW Conference,1998
 5. W.Xing and A.Gorbani,”Weighted PageRank Agorithm,” Proceedings of the Second Annual Conference on Communication Networks and Services Research,May 2004,pp. 305-314.
 6. N.Tyagi and S. Sharma,”Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page,”International Journal of Soft Computing and Engineerig(IJSCE),July 2012.
 7. G.Kumar, N. Duhan and A.K. Sharma,”Page Ranking Based on Number of Visits of Web Pages,”International Conference on Computer & Communication Technology(ICCCT, 2011,pp. 11-14.
 8. H. Dubey and Prof. B.N. Roy,”An Improved Page Rank Algorithm based on Optimized Normalization Technique,”International Journal of Computer Science and Information technologies(IJCSIT),2011,pp.2183-2188.
 9. N. Duhan, A.K. Sharma and Bhatia K.K., “Page Ranking Algorithm A Survey”, Proceeding of the International Conference on Advance Computing, pp. 128-135, 2009.
 10. D. K. Sharma and A. K. Sharma “, A Comparative Analysis o f the Page Ranking Algorithms”, International Journal of Computer Science and Engineering(IJCSE), pp. 2670-2776, 2010.
- C. Ding, X. He, H. Zha, P.Husbands and H. Simon “, Link Analysis: Hubs and Authorities on the World,” Technical Report: 47847, 2001.
- L. Page, S. Brin, R. Mtvani and T. Winogard “, The Page Ranking Citation Ranking: Bring Order to the Web,” Technical Report, Stanford Digital Libraries, SIDL-WP, 1999.