**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**
**IMPACT FACTOR: 7.056**

*IJCSMC, Vol. 10, Issue. 7, July 2021, pg.76 – 83*

# MACHINE LEARNING BASED SEARCH ENGINE WITH CRAWLING, INDEXING AND RANKING

**Anuradha T[1]; Tayyaba Nousheen[2]**
[1]Computer Science and Engineering Department and VTU University, India
[2]Computer Science and Engineering Department and VTU University, India
anuradhat26@gmail.com; tayyaban18@gmail.com
**DOI: 10.47760/ijcsmc.2021.v10i07.011**

***Abstract- The web is the heap and huge collection of wellspring of data. The Search Engine are used for retrieving the information from World Wide Web (WWW). Search Engines are helpful for searching user keywords and provide the accurate result in fraction of seconds. This paper proposed Machine Learning based search engine which will give more relevant user searches in the form of web pages. To display the user entered query search engine plays a major role of basic interface. Every site comprises of the heaps of site pages that are being made and sent on the server.***

***Key Terms:- Search Engine, XG Boost, Page Ranking, WWW, Web Crawler, indexing and ranking.***

## I. Introduction

Machine learning provides smart alternatives to analysis vast volume of data. By developing fast and effective algorithms and data driven models for real time processing of data, Machine learning can provide accurate result and analysis. Search Engine is a software program that is used for searching the WebPages based on user queries. Every site consists of heap of WebPages that are been created & saved on the server. The user needs to enter the set of keywords to get the desired result. Search Engine plays a vital role in the area of internet as it act as a interface between the search user query & displaying the result in the form of web pages. Here comes the actual need for search engines. Search engines provide you a simple interface to search user query and display the results in the form of the web address of the relevant web page. We are using machine learning in search engine for pattern detection that help in identifying the patterns and spam content.

 

The three main components of search engine

1. Web Crawler: Web Crawlers are used in collection of data about the websites and link them. We are using the information from WWW and collecting data from web and store the data in our database. And their purpose is to index the content of websites all across the internet so that websites can appear in search engine results.

2. Indexer: Data about the web pages are stored in the indexer as it is a database of a Search Engine. Indexer will arranges the keywords terms on each webpage and store them in the repository. When the user type the keyword in the search engine the relevant data is been searched in the indexer and it will provide the research result to the user.

3. Query Engine: It is a software that sites on the top of the server to provide the answer to the user searches. Query Engine help to reply the user query and shows displays the most probable outcome of the user searches, And they uses page ranking algorithm to rank the web pages and display it on the top.

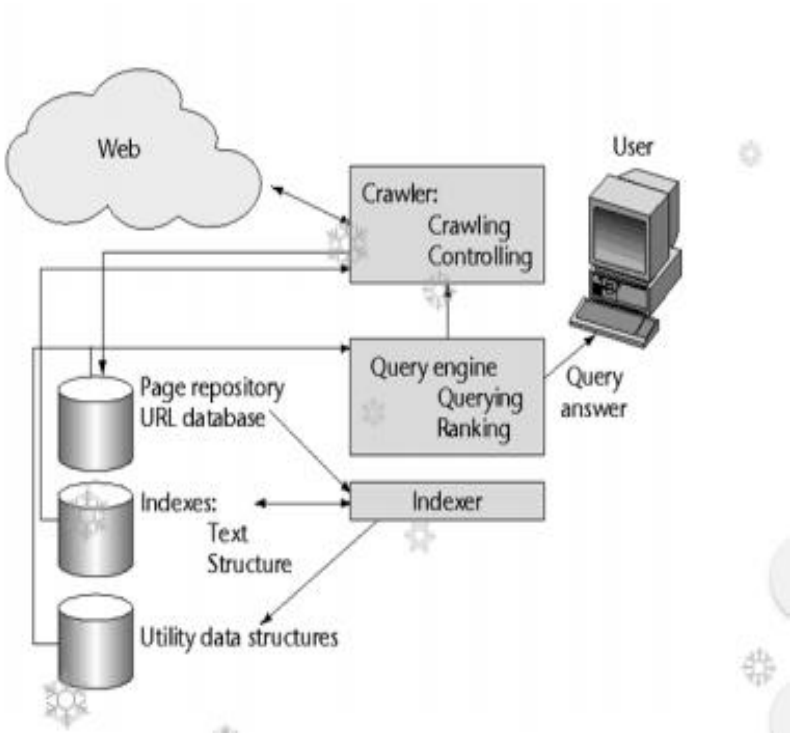


Figure 1. Block diagram of Search Engine.

## II. Literature Review

Numerous endeavors have been made by experts and researchers in the field of search engine. In [1],the authors discussed various types of search engines and they conclude the crawler based search engine is best among them and also Google uses it. It gives a user more relevant web address for user query. A Web crawler is a program that navigates the web by following the regularly changing, thick and circulated. Hyperlinked structure and from there on putting away downloaded pages in a vast database which is after indexed for productive execution of user queries. In [2], author concludes that major benefit of using keyword focused web crawler over traditional web crawler is that it works intelligently, efficiently.

The search engine uses a page ranking algorithm to give more relevant web page at the top of result, according to user need. It eases the searching method and user get required information very easily. Initially just an idea has been developed as user were facing problem in searching data so simple algorithm introduced which works on link structure, then further

 

modification came as the web is expanding. In [4], proposed a system which is based on a machine learning approach for web page filtering. The machine learning result is compared with traditional algorithm and found that machine learning result are more useful. The proposed approach is also effective for building a search engine. In [5], proposed the idea of page ranking algorithm which have been increasing rapidly over the years, it helped in retrieving the information required form the web and thereby filling the user needs and this paper gives the idea of various page ranking algorithms and is comparative study. In [6], the authors have compared various search engine in terms of performance, accuracy, images, average response time and results in showing the best search engine.The analysis of the search engine i.e. the integral parts and the steps by steps process how the search engine works and provide the accurate result for the user searched queries and it survey how to reduce the unwanted search results in the searching process is discussed in [7]. In [8], the machine learning techniques for search engine and studies concludes that how the directories are classified for storing the web pages in the search engine. The algorithms and functions for page ranking algorithm and stated that graph neural network is the best ranking algorithm which is used by the Google and shows the automated methods by most of the web pages for classification is observed in [9]. In [10], proposed the model of domain specific search engine is growing with popularity as they offer increased accuracy and extra functionality and used greatly in automate the creation and maintenance and describe new research in reinforcement learning, information extraction and text classification. The development of Meta search engine which provide for refining and classifying the search engine results and narrow down the results in a sequentially linked manner resulting in drastic reduction of web pages is studied in [11]. In [12], they have helped in knowing the machine learning techniques which deals with the creation and analysis of algorithm that can follow a reinforced methodology of learning. And shows the relations among the different concepts, required information is saved along with its strength. In [13], the authors have proposed the search engine cannot index every Web Pages, due to limited storage, Bw computation resources and the dynamic nature of the web. It cannot monitor continuously all parts of the web for changes. The combined techniques also include selection policy such as page rank, path ascending, focused crawling revisit policy such as freshness, age politeness. In [14], they proposed the study of the internal working procedure of the search engine how the search is process is done, crawling, indexing and ranking of the web page and usage of the ranking algorithm. In [15], they have analyzed the different tools and techniques to build the search engine such as navigational search queries, techniques such as algorithm for performance and ranking.

## III. Methodology

Search Engine allow researchers to enter search term, the engine then lists web pages on which information about the terms might be found. Search engine work by regularly sending out spider programs that search for newly appearing WebPages and then cataloging the content of these pages, And the database we are using be Full text Search which help in examining all of the words in every stored document as it tiers to match all the search criteria. Our aim is to build a search engine with increasing accuracy compare to different search engine. And to provide the most relevant web pages based on user queries. Here is the procedure to build a search engine:

A. Collection of data from World Wide Web: In this step, to collect data and information from internet we are using keyword based web crawler. Where www is the huge collection of data and known as information system where documents, web resources are identified by uniform resource locators

Algorithm:

Step 1: Start with URL.

Step 2: Queue Initialization.

Step 3: URL dequeue from queue

Step 4: To Download URL based web pages.

Step 5: To Extract URLs from downloaded web pages.

Step 6: The Extracted URL is inserted into queue.

Step 7: Return of step 1 until more relevant results have been achieved.

The above algorithm is a algorithm of collecting the data from the web, as we know web is a information system. Here we start with the url ex: https://example.com/ which is the hyperlink and are accessible by user over the internet.via web browser. And then the url is been initialized to the values of an existing list and it is removed from the data awaited

 

processing from a queue, then download the web page according to the url or copy the url of the web pages. And then the extracted url have been added to the queue and desired result have been obtained. And these process is done in fraction of seconds

B. Data cleaning process using natural language processing: Data cleaning is performed for preprocessing of data to remove the unnecessary data from the user search queries, and provide the needed results for the user so that the user will be help full in getting the desired data. NPL(natural language processing) is a field in ,machine learning with ability of a computer to understand, analysis, communicate with humans in their own language.

C. Comparison the page ranking algorithm: Here we are comparing two page ranking algorithm which gives more accuracy and efficiency that we are going to use in our search engine.

Table I. Comparison between Page rank and Efficiency weighted page rank

| Requirements | Page Rank (PR) | Efficiency Weighted Page Rank (WPR) |
|---|---|---|
| Working | The algorithm calculates the no. of visit and ranks it accordingly. | The algorithm calculates the in and out bound link of the importance WebPages. |
| Input | Inbound links | Inbound and outbound links |
| Algorithm Complexity | O(log N) | < O(log N) |
| Quality of Results | Good | More than Page Rank |
| Efficiency | Medium | High |

So the algorithm which gives highest accuracy and efficiency is WPR. Here we are comparing the two algorithms based on the existing research, working ranking is been calculated based on the number of visits, where as inbound links are the links which are link from the some other website to the web resources, outbound links are the links which will direct you to the specific web page.

D. Merging the selected page ranking algorithm with machine learning techniques: After selecting the most relevant Page Rank algorithm, that algorithm is considered as input for machine learning algorithm. The output of this algorithm will give the web address of most relevant web page to the user search queries. Implementation of query engine to display the user query with efficient results. It is the last step, In this implemented Query engine will takes the input from the user in a form of quires and display the results in the form of relevant web page.

Implementation of web platform to displaying the user query with efficient results: The final step is to built a web platform for the user so that they can search the query and obtain a efficient result.

For speed and accuracy we are using the algorithms that use machine learning techniques. Here we are going to compare some algorithm they are

1) Support Vector Machine
2) Artificial Neural Network
3) XG Boost

 

Table II . Accuracy of different algorithm

| No. | Algorithm | Accuracy |
|---|---|---|
| 1 | SVM | 89.58 |
| 2 | ANN | 91.36 |
| 3 | XGBoost | 92.58 |

The following above table shows the accuracy of each algorithm out of all the algorithm , XGBoost will provide the highest accuracy.

Accuracy is calculated by using the formula= $\frac{\text{number of documents classified correctly}}{\text{Total number of documents}}$

SVM (support vector machine): It is the algorithm in machine learning to solve the linear and non linear problems and work with many practical problems. And practically accuracy is measured as 89.58%.

ANN (Artificial neural network): In this algorithm is to make an attempt to simulate the network of neurons that make up a human brain so that able to learn things and make decision like humans. And practically accuracy is measured as 91.36%.

Xgboost: It is the algorithm that has recently been applied in machine learning for structured data, and designed for speed and accuracy. And practically accuracy is measured as 92.58%.

Step1: Start with the browser
Step2: Enter the input
Step3: Load all the libraries
- Library (xgboost)
- Library (readr)
- Library (string)
- Library (caret)
- Library (car)

Step4: Load the data in the database
Step5: Data cleaning for the user input
Step6: Tune and run the web browser
Step7: Provide the efficient result.
Step8: if not, repeat step 2
Step9: stop.

In this algorithm we have to load all the libraries, dataset and clean the data according to the user input and then click on the search it will run the model and provide the efficient result.

# IV. Working Algorithm

Here is the algorithm for searching the result,

Step 1: Start with the browser
Step 2: Input query form user
Step 3: Permit auto suggestion operation
Step 4: Perform auto correction operation
Step 5: Match the query with database table contents records to find the matching results
Step 6: store all matched records in variable
Step 7: Sort the results as per ranking

 

Step 8: Display the results with time taken to display the search results
Step 9: If accurate result is not found repeat step2
Step 10:If accurate result are found then stop the process
Step 11: Stop.

Here we have to start with the web browser, then enter the user input and the web platform will help the user in auto suggestion and auto correction of the input. Match the input with the stored database and sort the result according to ranking and display the efficient result according to the result.

# V. Experimental Results and Discussion

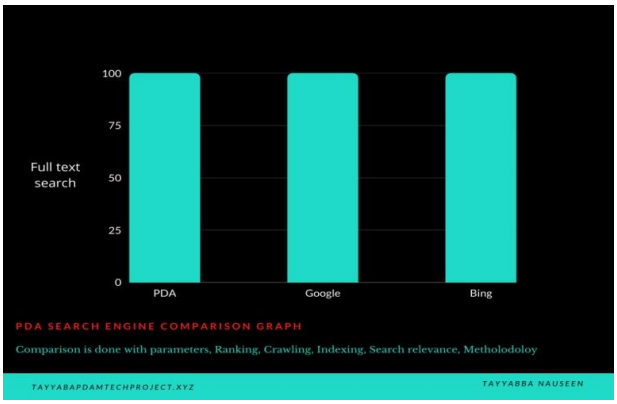


Fig 2: Full Text Search

All three search engines use full text search and are same. All full text are same in performance is same except depends on the speed on the platform used.Full text search, is a more advanced way for searching a db. It quickly finds a result from the db and provides to the end user. As full text is a technique for searching a computer stored document. A full text query returns any documents that contains at least one match and eases to search the content from the database.

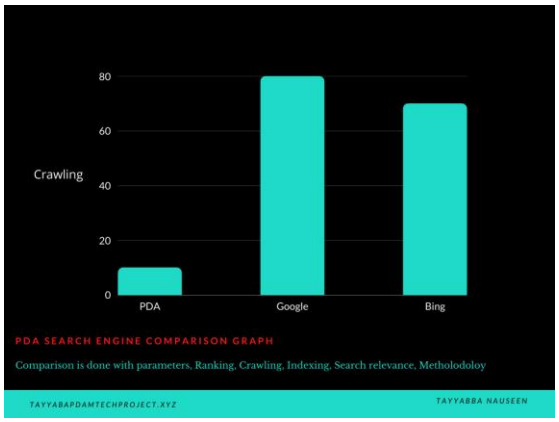


Fig 3: Crawling is the process used by search engine web crawlers (bots or spiders) to visit and download a page and extract its links in order to discover additional pages

1. PDA: 1%-2% of the web
2. Google: 80% of the web (google bot crawls almost all websites on the web except dark web websites)
3. Bing: 70% of the web (google bot crawls almost all websites on the web except dark web websites)

 

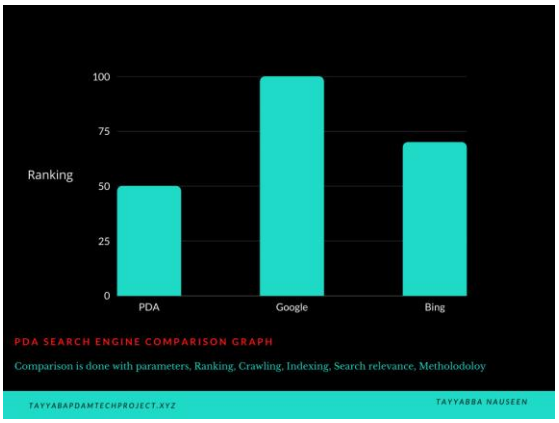


Fig4: Ranking

Ranking is a process in which the web pages have been shown according to the number of user visited the web page. The first displayed web page is the page with most of the visited.

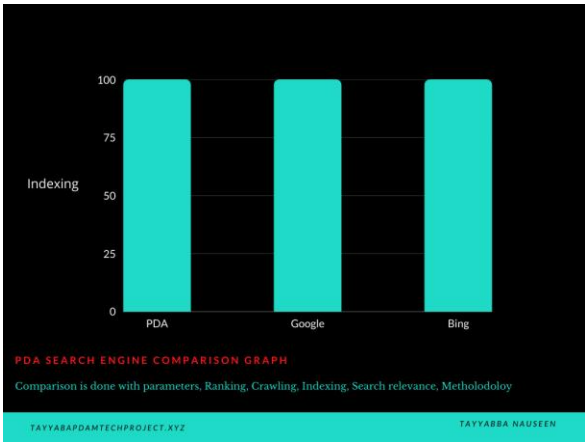


Fig5: Indexing: (Graph → any website can be indexed)

Indexing is a process in which response to the information query has been done fast. As it stored in the database and has been indexed, any website can be indexed which has permission in robot.txt file on server and can be added to the database.

## Conclusion:

Search engine is very useful for finding out more relevant URL for given keyword. Due to this, user time is reduced for searching the relevant web page. For this, Accuracy is very important factor. From the above observation, it can be concluded that Xgboost is a best in terms of accuracy than SVM and ANN. Thus, Search engine built using Xgboost and Page Rank algorithm will give better accuracy. The amount of information available on the Internet continues to grow exponentially. As this trend continues, we argue that not only will the public need powerful tools to help them sort through this information, but the creators of these tools will need intelligent techniques to help them build and maintain these services. The amount of information available on the Internet continues to grow exponentially. As this trend continues, we argue that, not only will the public need powerful tools to help them sort though this information, but the creators of these tools will need intelligent techniques to help them build and maintain these tools.

 

# References:


**[1]** Manika Dutta, K. L. Bansal, "A Review Paper on Various Search Engines (Google, Yahoo, Altavista, Ask and Bing)", International Journal on Recent and Innovation Trends in Computing and Communication, 2016.pp.190-196

**[2]** Gunjan H. Agre, Nikita V.Mahajan, "Keyword Focused Web Crawler", International Conference on Electronic and Communication Systems, IEEE, 2015. pp.463-465

**[3]** Tuhena Sen, Dev Kumar Chaudhary , "Contrastive Study of Simple Page Rank, HITS and Weighted Page Rank Algorithms: Review", Inter- national Conference on Cloud Computing, Data Science & Engineering, IEEE, 2017. pp.67-75

**[4]** Michael Chau, Hsinchun Chen, "A machine learning approach to web page filtering using content and structure analysis", Decision Support Systems 44 (2008) 482–494,scienceDirect,2008. ppm276-280

**[5]** Rushikesh karwa , Vikas Honmane , "Building Search Engine Using Machine learning techniques", International Conference on Intelligent Computing and control systems , 2019. pp.1061-1064

**[6]** Joseph Edosomwan, Taiwo O, "Comparative analysis of some search engines" College of Education, Ekiadolor-Benin, Edo State, Nigeria. pp.1-4

**[7]** R. Rubini1 ,R. Manicka Chezian, "An Analysis on Search Engines: Techniques and Tools" Research Scholar, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India 1. pp.7860-7864

**[8]** Neenu Ann Sunny, "Machine Learning in Search Engines", Assistant Teacher Saintgits College of Applied Science. pp.155-161

**[9]** Sweah Liang Yong, Markus Hagenbuchner, "Ranking Web Pages using Machine Learning Approaches", University of Wollongong. pp.1-4

**[10]** Andrew McCallumzy ,Kamal Nigamy, knigam Renniey ,Kristie Seymorey, "Building Domain-Specific Search Engines with Machine Learning Techniques", Just Research 4616 Henry Street & School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213. pp.662-667

**[11]** Vishwas Raval and Padam Kumar, "SEReleC (Search Engine Result Refinement and Classification) - A Meta Search Engine based on Combinatorial Search and Search Keyword based Link Classification", M Tech Scholar, Professor & Head, Electronics & Computer Engineering Department Indian Institute of Technology Roorkee, Uttarakhand, India. pp.1-5

**[12]** Pratiba D , Shobha G , Samrudh J , "Personalized Web Search using Machine Learning Technique". pp.118-122

**[13]** Suyash Gupta, KrishanDev Mishra, Prerna Singh, "Effective Searching Policies for Web Crawler". pp.3137-3139

**[14]** S. Prabha, K. Duraiswamy, J. Indhumathi, "Comparative Analysis of Different Page Ranking Algorithms", International Journal of Computer and Information Engineering, 2014.pp.1546-1554

**[15]** Dilip Kumar Sharma, A. K. Sharma, "A Comparative Analysis of Web Page Ranking Algorithms", International Journal on Computer Science and Engineering, 2010. pp.2670-2677


 