Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019)
IEEE Xplore Part Number: CFP19K34-ART; ISBN: 978-1-5386-8113-8

Building Search Engine Using Machine Learning Technique

1st Rushikesh Karwa
Department of Computer Science and Engineering
Walchand College of Engineering
Sangli(MS), India
rushikeshkarwa55@gmail.com

2nd Vikas Honmane
Department of Computer Science and Engineering
Walchand College of Engineering
Sangli(MS), India
vhonmane@gmail.com

***Abstract*—The web is the huge and most extravagant well-spring of data. To recover the information from World Wide Web, Search Engines are commonly utilized. Search engines provide a simple interface for searching for user query and displaying results in the form of the web address of the relevant web page, but using traditional search engines has become very challenging to obtain suitable information. This paper proposed search engine using Machine Learning technique that will give more relevant web pages at top for user queries.**

***Index Terms*—World Wide Web, Search Engine, PageRank, Machine Learning.**

I. Introduction

World Wide Web is actually a web of individual systems and servers which are connected with different technology and methods. Every site comprises of the heaps of site pages that are being made and sent on the server. So if a user needs something, then he or she needs to type a keyword. Keyword is a set of words extracted from user search input. Search input given by user may be syntactically incorrect. 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.

Figure 1 focuses on three main components of search engine.

1) Web crawler
   Web crawlers help in collecting data about a website and the links related to them. We are only using web crawler for collecting data and information from WWW and store it to our database.
2) Indexer
   Indexer which arranges each term on each web page and stores the subsequent list of terms in a tremendous repository.
3) Query Engine
   It is mainly used to reply the user's keyword and show the effective outcome for their keyword. In query engine, Page ranking algorithm ranks the URL by using different algorithms in the query engine.

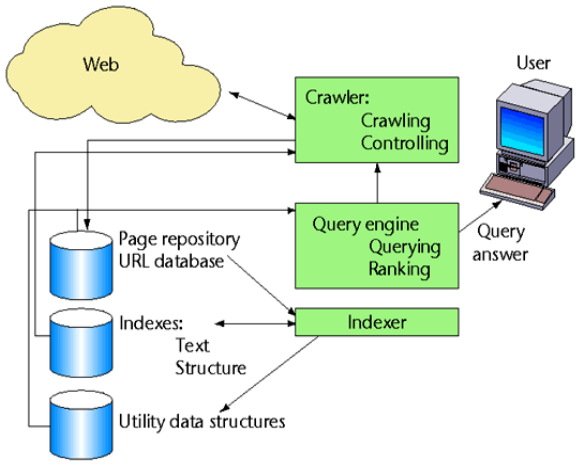

Fig. 1. Block Diagram of Search Engine [1]

This paper utilizes Machine Learning Techniques to discover the utmost suitable web address for the given keyword. The output of PageRank algorithm is given as input to machine learning algorithm.

The section II discusses the related work in search engine and PageRank algorithm. In section III Objective is explained. Section IV deals with proposed system which is based on machine learning technique and section V contains the conclusion.

II. Literature Review

Numerous endeavors have been made by data experts and researchers in the field of search engine. Dutta and Bansal [1] discuss various type of search engine and they conclude the crawler based search engine is best among them and also Google uses it. It give 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 conclude 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 ease 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 also expanding so weighted PageRank and HITS came into the scenario. In [3], author compare various PageRank algorithm and among all, Weighted PageRank algorithm is best suited for our system. Michael Chau and Hsinchun Chen [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.

III. Objective

To build a search engine which gives web address of the most relevant web page at the top of the search result, according to user queries. The main focus of our system is to build a search engine using machine learning technique for increasing accuracy compare to available search engine.

IV. Methodology

To build a search engine which gives web address of the most relevant web page at the top of the search result, according to user queries. The main focus of our system is to build a search engine using machine learning technique for increasing accuracy compare to available search engine.

Following is the step by step procedure for building the search engine:

1) Collect data from WWW using web crawler.
2) Perform data cleaning using NLP.
3) Study and compare the existing page ranking algorithm.
4) Merge the selected page rank algorithm with current technologies in machine learning.
5) Implement query engine to display the efficient results for user query.

A. Collect data from WWW using web crawler

In this step, we are using keyword based web crawler to collect data and information from internet. It begins its working utilizing seed URL. Subsequent to visiting the website page of seed URL and concentrates every one of the hyperlinks present in that site page and store the extracted hyperlinks to the queue and exact the data from all web pages. Finally filter out the URL which is relevant for particular keywords.

Algorithm steps:

Step 1: Start with seed URL.
Step 2: Initialize queue (q).
Step 3: Dequeue URL's from queue (q).
Step 4: Downloads web page related with this URL.
Step 5: Extract all URLs from downloaded web pages
Step 6: Insert extracted URL into queue (q).
Step 7: Goto step 1 until more relevant results are achieved.

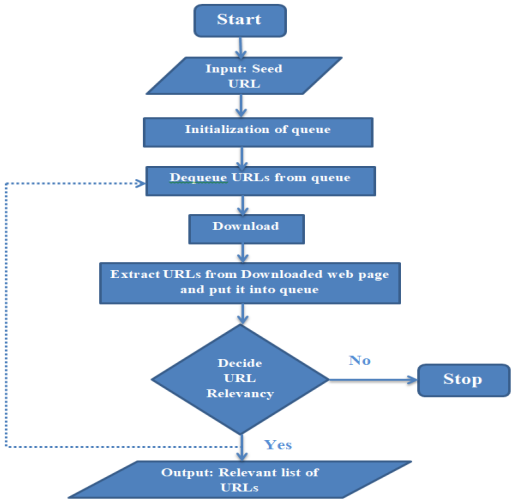

Fig. 2. Flowchart for keyword focused web crawler [2]

B. Perform data cleaning using NLP

In this step, data cleaning is performed to preprocess the data using NLP steps so that unnecessary data is removed. After collecting data from WWW using web crawler, there is need to perform data cleaning using NLP.

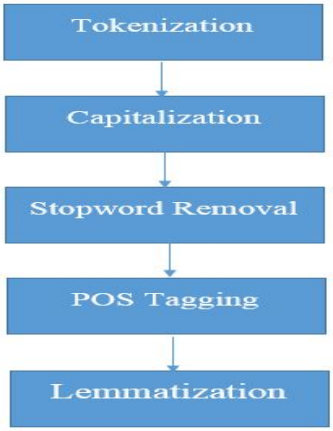

Fig. 3. NLP steps for data cleaning

Figure 3 shows the step by step procedure for data cleaning using NLP.

1) Tokenization = Tokenization depicts divide web page passages into phrases, or phrases into individual terms.
2) Capitalization = The most common approach is to reduce all web page data to lower case for simplicity.

3) Removing the stopword = Web page consists of many words which are mainly used for connecting parts of the sentence rather than showing important information. Here is need to remove such words.
4) Parts of Speech tagging = It will split the sentence into token and give significance for each token which is also applies to user queries in search engines to find out exactly what user required?
5) Lemmatization = Lemmatization is where words are diminished to a root by evacuating affectation through dropping pointless characters.

C. Study and compare the existing page ranking algorithm

TABLE I
COMPARISON BETWEEN PR, WPR AND HITS [7]

Criteria	PageRank (PR)	Weighted PageRank (WPR)	HITS
Working	This algorithm calculates the page score at the time the pages are indexed.	Web page weight is calculated based on inbound and outbound links of importance web page.	It calculates hub and authority score for each web page.
Input Parameter	Incoming links	Incoming and outgoing links	Content, incoming and outgoing links
Algorithm Complexity	O(log N)	< O(log N)	< O(log N)
Quality of Results	Good	More than PageRank	Less than PageRank
Efficiency	Medium	High	Low

Among all, the Weighted PageRank algorithm is best suited for system because it gives more accuracy and efficiency comparable to other (see table-1).

D. Merge the selected page rank algorithm with current technologies in machine learning

After selecting and implementing the best suited PageRank algorithm. In this step, topmost output of pagerank algorithm is considered as input for machine learning algorithm. The output of machine learning algorithm is given to the user as a web address of relevant web page based on user queries.

For implementing the machine learning algorithm to find out the most relevant web page based on user queries, we are dividing the web feature into three parts:

1) Page content
2) Page content of Neighbors
3) Link analysis

1) Page Content
   Instead of a word vector, content of each web page is represented by following feature scores.

   Five feature scores are defined :

   a) Title (w) = Number of words within the title of web page w found associated with given query.
   b) TFIDF (w) = Sum of TFIDF of the words in web page w found associated with given query.
   c) URL (w) = Number of words within the URL of web page w found associated with given query.
   d) Heading (w) = Number of words within the heading of web page w found related to given query.
   e) Anchor (w) = Number of words found in the anchor text describing web page w related to the particular query.

2) Page content of neighbors
   In this methodology, three sorts of neighbors were considered:inbound, outbound, and sibling. For any web page w, inbound neighbors are the collection of all web pages that have a hyperlink to w. Outbound neighbors are the collection of all web pages whose hyperlinks are found in w. Sibling pages are set of all web pages that are pointed any of the parents of w.

   Six feature scores are defined :

   a) InTitle (w) = Mean (number of words within the title of web page m found associated with given query) for all inbound web pages m of w.
   b) InTFIDF (w) = Mean (sum of TFIDF of the words in page m found associated with given query) for all inbound web pages m of w.
   c) OutTitle (w) = Mean (number of words within the title of web page n found associated with given query) for all outbound web pages n of w.
   d) OutTFIDF (w) = Mean (sum of TFIDF of the words in web page n found associated with given query) for all outbound pages n of w.
   e) SiblingTitle (w) = Mean (number of words within the title of web page q found associated with given query) for all sibling web pages q of w.
   f) SiblingTFIDF (w) = Mean (sum of TFIDF of the words in web page q found associated with given query) for all sibling web pages q of w.

3) Link analysis
   Link are nothing but hyperlinks from one page to another that are also very useful for deciding the relevancy and quality of the web page.

   Three feature scores were defined :

   a) PageRank (w) = PageRank of web page w.
   b) Inlinks (w) = Count of inbound links pointing to w.
   c) Outlinks (w) = Count of outbound links from w.

All of above web features are considered as an input feature for ANN,SVM and Xgboost and the one who give more accuracy is merged with the selected pagerank algorithm to give the URL of relevant web pages for user queries.

E. Implement query engine to display the efficient results for user query

At last, implement the Query engine which takes the input from the user in a form of query and display the efficient result

for their query. It will display the web address of relevant pages based on the output of machine learning algorithm.

V. Experimental result

Following is a list of algorithm which is implemented. The algorithm which give more accuracy is used which PageRank algorithm.

1) Support Vector Machine
2) Artificial Neural Network
3) XGBoost

1) Support Vector Machine

Because of its exceptional performance, a SVM was also used to allow a better approach. It used the same set of feature scores to perform classification.
Dataset is not linearly separable so we are using non-linear SVM. Rbf, poly and sigmoid are type of non-linear kernel. The above 14 feature are selected as a input for SVM model and based on that feature, SVM tried to predict, whether each web page in the testing set was relevant to the given query or not. The results were stored and used for performance evaluation.

2) Artificial Neural Network

A neural network consist of three layers, namely input layer, hidden layer, and output layer. The neural network's input layer consisted of 14 nodes corresponding to each web page's 14 feature scores.
Only one output node is required in output layer for determining relevancy of a web page. The number of nodes was set to 7 in the hidden layer. These parameters are set using a grid search based on some initial experimentation. The entire process has been repeated 150 times and the batch size is set to 10. The results were stored and used for performance evaluation.

3) XGBoost
It is a type of Boosting based ensemble learning. It uses gradient boosted decision trees for improving accuracy and speed.
The input feature consist of same 14 features and we are using gbtree based booster. The number of classifier are set to 50 and max_depth size is set to 4. These parameters are set using a parameter turning and cross validation based on some initial experimentation.

A. Performance

1) Accuracy

Firstly, we create a dataset of 540 records in which the dependent variable is the relevancy of URL which is 0 or 1. 1 indicate relevant URL and 0 indicate irrelevant URL. Dataset is divided into training and testing dataset. Out of 540 records, 378 record are used for training purpose and 162 are used for testing purpose. The accuracy is calculated using following formula.

$$accuracy = \frac{number\ of\ documents\ correctly\ classified}{total\ \ number\ of\ documents}$$

The following table show the accuracy of each algorithm. Out of all, XGBoost have more accuracy.

TABLE II
Accuracy of different algorithm

No.	Algorithm	Accuracy
1	SVM	89.50
2	ANN	91.35
3	XGBoost	92.59

VI. 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 PageRank algorithm will give better accuracy.

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.
[2] Gunjan H. Agre, Nikita V.Mahajan, "Keyword Focused Web Crawler", International Conference on Electronic and Communication Systems, IEEE, 2015.
[3] Tuhena Sen, Dev Kumar Chaudhary, "Contrastive Study of Simple PageRank, HITS and Weighted PageRank Algorithms: Review", International Conference on Cloud Computing, Data Science & Engineering, IEEE, 2017.
[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.
[5] Taruna Kumari, Ashlesha Gupta, Ashutosh Dixit, "Comparative Study of Page Rank and Weighted Page Rank Algorithm", International Journal of Innovative Research in Computer and Communication Engineering, February 2014.
[6] K. R. Srinath, "Page Ranking Algorithms – A Comparison", International Research Journal of Engineering and Technology (IRJET), Dec-2017.
[7] S. Prabha, K. Duraiswamy, J. Indhumathi, "Comparative Analysis of Different Page Ranking Algorithms", International Journal of Computer and Information Engineering, 2014.
[8] Dilip Kumar Sharma, A. K. Sharma, "A Comparative Analysis of Web Page Ranking Algorithms", International Journal on Computer Science and Engineering, 2010.
[9] Vijay Chauhan, Arunima Jaiswal, Junaid Khalid Khan, "Web Page Ranking Using Machine Learning Approach", International Conference on Advanced Computing Communication Technologies, 2015.
[10] Amanjot Kaur Sandhu, Tiewei s. Liu., "Wikipedia Search Engine: Interactive Information Retrieval Interface Design", International Conference on Industrial and Information Systems, 2014.
[11] Neha Sharma, Rashi Agarwal, Narendra Kohli, "Review of features and machine learning techniques for web searching", International Conference on Advanced Computing Communication Technologies, 2016.
[12] Sweah Liang Yong, Markus Hagenbuchner, Ah Chung Tsoi, "Ranking Web Pages using Machine Learning Approaches", International Conference on Web Intelligence and Intelligent Agent Technology, 2008.
[13] B. Jaganathan, Kalyani Desikan,"Weighted Page Rank Algorithm based on In-Out Weight of Webpages", Indian Journal of Science and Technology, Dec-2015.