Building Search Engine Using Machine Learning Technique

Ch.Venkata Ramana[1], G. Meghana[2], M. Navya Sai[3,] A. Prasad[4], V. Mohanarao[5]
[1,2,3,4,5]*Pragati Engineering College*

Abstract **- Building Search Engine using Machine Learning Technique The web is the huge and most extravagant wellspring of data. To recover the information from the 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 a search engine using Machine Learning technique that will give more relevant web pages at top for user queries.**

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 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 a user may be syntactically incorrect. Here comes the actual need for search engines. Search engines provide you a simple interface to search user queries and display the results. • Web crawlers help in collecting data about a website and the links related to them. We are only using web crawlers for collecting data and information from WWW and storing it in our database. • Indexer which arranges each term on each web page and stores the subsequent list of terms in a tremendous repository. • Query Engine is mainly used to reply to the user's keyword and show the effective outcome for their keyword. In the query engine, the Page ranking algorithm ranks the URL by using different algorithms in the query engine. • This paper utilizes Machine Learning Techniques to discover the utmost suitable web address for the given keyword. The output of the Page Rank algorithm is given as input to the machine learning algorithm.

1.1 MOTIVATION

As in today's internet world, people are mostly based on search engines to search what they are looking for in the internet. • The web is the huge and most extravagant well spring of data. To retrieve the information from the World Wide Web, Search Engines are commonly utilized.

1.2 PROBLEM DEFINTION: The project we have built is used to provide the faster retrieval of information using search engines that are implemented by using machine learning algorithms. It provides 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

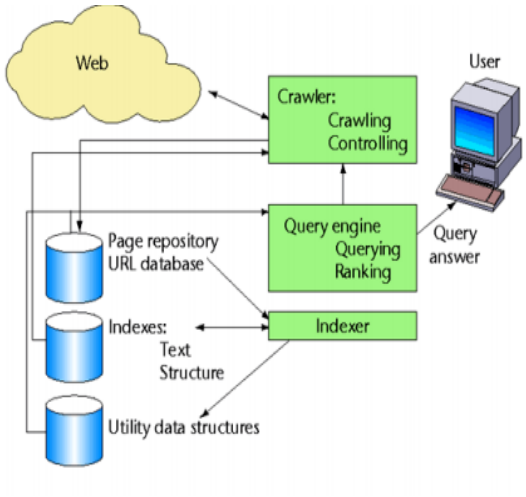
1.3 OBJECTIVE OF PROJECT: To build a search engine which gives web address of the most relevant web page at the top of the search result, according to user queries. The main focus of our system is to build a search engine using machine learning technique for increasing accuracy compare to available search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

EXISTING SYSTEM • Information retrieval is to retrieve the information resources that we are interested in or extract whatever information we need. • Information Retrieval (IR) may deal with the organization, storage, retrieval and evaluation of information from documents, particularly textual information. • But we cannot give the ranks to those documents.

DISADVANTAGES OF EXISTING SYSTEM: • Information retrieval will be very difficult in large numbers of texts in a document. • Difficult to identify the important concepts or topic in a collection of

documents. • The explicit rankings are always difficult to obtain or even not available in many documents.

II. SYSTEM DESIGN

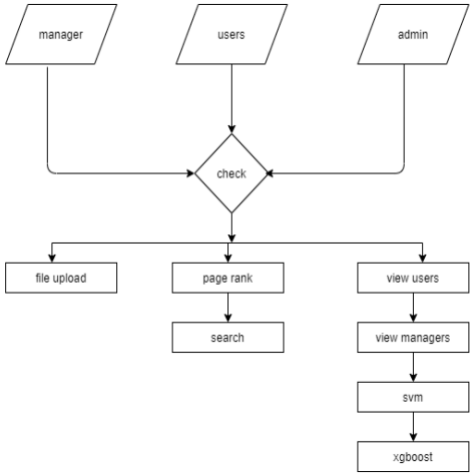


1. System Architecture

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. The figure 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.

DATA FLOW DIAGRAM OF PROJECT

The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system. • The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, an external entity that interacts with the system and the information flows in the system



PROPOSED SYSTEM • The proposed search engine is very useful for finding out more relevant URLs for given keywords. • Anyone can easily identify the important documents in a collection of documents and retrieve the related data. • It proposes a novel model, named LDA (Linear Discriminant Analysis), easy for clustering the related documents based on that ranking

ADVANTAGES OF PROPOSED SYSTEM:

- We will build a search engine which gives the 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 to discover the utmost suitable web address for the given keyword by using machine learning techniques for increasing accuracy compared to available search engines.

REQUIREMENT ANALYSIS

The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

REQUIREMENT SPECIFICATION

Functional Requirements:

Graphical User interface with the User. Software Requirements:

For developing the application the following are the Software Requirements: • Python • Django Operating

Systems supported • Windows 7 • Windows XP • Windows 8 Technologies and Languages used to Develop

IMPLEMENTATION

We have used three algorithms in our project. They are:

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

SUPPORT VECTOR MACHINE:

SVM is treated as 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 nonlinear SVM. Rbf, poly and sigmoid are type of nonlinear 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.

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. 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

INPUT AND OUTPUT DESIGN INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things: • What data should be given as input? • How the data should be arranged or coded? • The dialog to guide the operating personnel in providing input. • Methods for preparing input validations and steps to follow when error occur. OBJECTIVES 1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system. 2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities. 3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

OUTPUT DESIGN

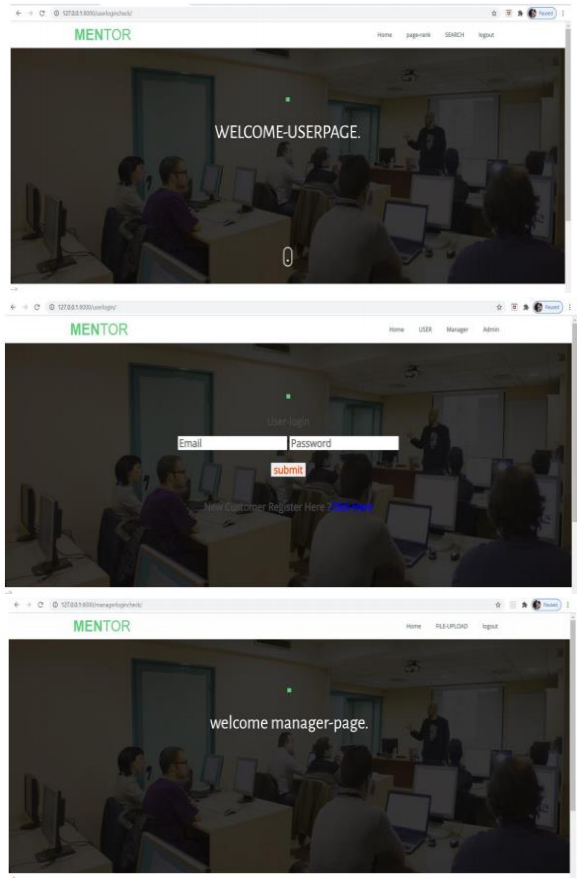
A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making. 1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed

to meet the requirements. 2. Select methods for presenting information. 3. Create document, report, or other formats that contain information produced by the system.

SOFTWARE ENVIRONMENT

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. An interpreted language, Python has a design philosophy that emphasizes code readability (notably using whitespace indentation to delimit code

OUTPUT SCREENS



III.CONCLUSION

Search engines are very useful for finding out more relevant URLs for given keywords. Due to this, user time is reduced for searching the relevant web page. For this, Accuracy is a very important factor. From the above observation, it can be concluded that XGBoost is better in terms of accuracy than SVM and ANN. Thus, Search engines built using XGBoost and PageRank algorithms will give better accuracy.

ACKNOWLEDGMENT

We would like to express our gratitude to all the people behind the screen who helped us to transform an idea into a real application. We would like to express our heart-felt gratitude to our parents without whom We would not have been privileged to achieve and fulfill our dreams. We are grateful to our principal, Dr. S SAMBHU PRASAD who most ably run the institution and has had the major hand in enabling me to do our project. We profoundly thank Dr. D SIRISHA, Head of the Department of Information Technology who has been an excellent guide and also a great source of inspiration to our work. We would like to thank our internal guide Mr. CH VENKATA RAMANA for his technical guidance, constant encouragement and support in carrying out our project. The satisfaction and euphoria that accompany the successful completion of the task would be great but incomplete without the mention of the people who made it possible with their constant guidance and encouragement crowns all the efforts with success. In this context, We would like to thank all the other staff members, both teaching and non-teaching, who have extended their timely help and eased our task. PROJECT ASSOCIATES G. MEGHANA 17A31A1207 M. NAVYA SAI 17A31A1215 A. PRASAD 17A31A1241 V. MOHANA RA 17A31A125

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), Dec2017.
[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