Building Search Engine Using Machine Learning Technique

V Krishna reddy[1], S. Sreeja[2], R.DurgaVamsi[3], P. Sanjay[4]
[1,2,3,4] *Information Technology, JB Institute of Engineering and Technology, Hyderabad, India*

***Abstract*— 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.**

1.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.

1) Web crawler 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.

2) 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 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.

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

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

2. LITERATURE SURVEY

1) Weighted page rank algorithm based on in-out weight of web pages

AUTHORS: Kalyani Desikan, B. Jaganathan.

In its classical formulation, the well known page rank algorithm ranks web pages only based on in-links between web pages. We propose a new in-out weight based page rank algorithm. In this paper, we have introduced a new weight matrix based on both the in-links and out-links between web pages to compute the page ranks. We have illustrated the working of our algorithm using a web graph. We notice that the page rank values of the web pages computed using the original page rank algorithm and our proposed algorithm are comparable. Moreover, our algorithm is found to be efficient with respect to the time taken to compute the page rank values.

2) Web Page Ranking Using Machine Learning Approach

AUTHORS: Junaid Khan, Arunima Jaiswal.

One of the key components which ensures the acceptance of web search service is the web page ranker - a component which is said to have been the main contributing factor to the early successes of Google. It is well established that a machine learning method such as the Graph Neural Network (GNN) is able to learn and estimate Google's page ranking algorithm. This paper shows that the GNN can successfully learn many other web page ranking

methods e.g. Trust Rank, HITS and OPIC. Experimental results show that GNN may be suitable to learn any arbitrary web page ranking scheme, and hence, may be more flexible than any other existing web page ranking scheme. The significance of this observation lies in the fact that it is possible to learn ranking schemes for which no algorithmic solution exists or is known.

3) Review of features and machine learning techniques for web searching.
AUTHORS: Neha Sharm , Narendra Kohli
As the amount of information is growing rapidly on world wide web, it has become very difficult to get relevant information using traditional search engines within a stipulated time. The main reasons for irrelevant search results are the lack of understanding of user's search intention or user's preferences, keyword based searching, short queries. In this paper, we will study different features that are used in information retrieval. We will also discuss various machine learning techniques that are helpful in deciding the relevance of web page to user. We have done classification on the basis of features. In the end we will compare different techniques and their pros and cons are also discussed.

3. MODULES

Manager:
Manager information and task descriptions for the entire experiment. Manager can upload the file into the database. we can upload the file with file type and name of the file and also particular url to the file to get the information about the file.

User:
User information and task descriptions for the entire experiment. User after login into the session he will get two options. he can search the whatever particular url or information. we can search the particular file and also we can get the weight and rank of the file by using the tfidf concept.

Admin:
Admin will give authority to managers and users. In order to facilitate activate the managers and activate the users. the admin can see the details of all users and managers. Admin can get the accuracy results of svm and xgboost algorithms.

Machine learning:
Machine learning refers to the computer's acquisition of a kind of ability to make predictive judgments and make the best decisions by analyzing and learning a large number of existing data. The representation algorithms include deep learning, artificial neural networks, decision trees, enhancement algorithms and so on. The key way for computers to acquire artificial intelligence is machine learning. Nowadays, machine learning plays an important role in various fields of artificial intelligence. Whether in aspects of internet search, biometric identification, auto driving, Mars robot, or in American presidential election, military decision assistants and so on, basically, as long as there is a need for data analysis, machine learning can be used to play a role.

4. 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.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

5. ALGORITHM

PAGE RANK ALGORITHM

The PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. It is assumed in several research papers that the distribution is evenly divided among all documents in the collection at the beginning of the computational process. The PageRank computations require several passes, called "iterations", through the collection to adjust approximate PageRank values to more closely reflect the theoretical true value.

6. 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.

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 ofInnovative 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 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