IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

CREATING A SEARCH ENGINE USING MACHINE LEARNING METHODS

R. N. Behera[1], A.K.Palit[2], S. K. Das[3]
[1]Assistant Professor, Dept of CSE, GITAM, Bhubaneswar - 752054.
[2] Assistant Professor, Dept of CSE, GITAM, Bhubaneswar - 752054.
[3] Assistant Professor, Dept of CSE, GITAM, Bhubaneswar - 752054.

ABSTRACT

The internet is a vast source of information. Search engines are commonly used to find information on the web. They offer a simple way to search for what users are looking for, but traditional search engines often struggle to provide the most relevant information. This paper suggests using machine learning techniques to build a search engine that will show the most relevant web pages at the top of search results 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.

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 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.

II. SYSTEM ANALYSIS

EXISTING SYSTEM:

To create a new Programmable Search Engine, all you have to do is choose which sites to search and give your search engine a name. In the Sites to search box, type one or more

sites you want to include in the search results. You can include any sites on the web, even sites you don't own. By our calculations, for mid to large size retailer to build its own high-quality, Solr-based site search engine would take 30 to 40 engineers as long as two years.140+ Search Engines and Directories. Search, the holy grail that pushed Google into global Internet domination, is still coveted by many. The fact that most users don't even consider switching Google for anything else doesn't mean that there's no innovation going on in the field of search.

DISADVANTAGES OF SEARCH ENGINE

- Sometimes the search engine takes too much time to display relevant, valuable, and informative content.
- Search engines, especially Google, frequently update their algorithm, and it is very difficult to find the algorithm in which Google runs.

PROPOSED SYSTEM:

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.

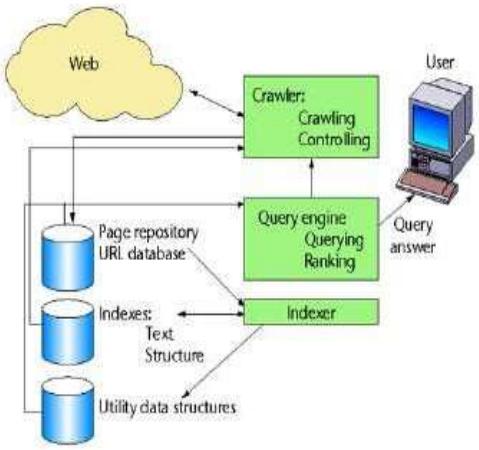
4) Merge the selected page rank algorithm with current technologies in machine learning.

5) Implement query engine to display the efficient results for user query.

ADVANTAGES:

- Time Savings. A **search engine** saves you time in two ways: by eliminating the need to find information manually, and by performing **searches** at high speeds.
- Relevance. When a **search engine** scans a website, it scores the content for relevance to particular **search** words.
- Free Access.
- Comprehensive.
- Advanced **Search**.

SYSTEM ARCHITECTURE:



III. IMPLEMENTATION

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 tf idf 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 xg boost 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.

IV. CONCLUSION

Search engines help find more relevant web pages for specific keywords, saving users time. Accuracy is crucial for this. Based on our observations, XGBoost performs better than SVM and ANN in terms of accuracy. Therefore, search engines using XGBoost and PageRank algorithms will provide 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), 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