Turkish Online Journal of Qualitative Inquiry
Volume 13, Issue 1, January 2022: 917-928

Machine Learning Techniques for Search Engine Development

Mr. Ayyapu Sai Sreenivas, UG Scholar**,**VIT University(IT),Vellore,Tamil Nadu, India.
Mr. Yashwanth Penugonda,VIT University(IT),Vellore, Tamil Nadu, India.
Mr. Achyuta Sai Nikhil Digital Marketing Analyst-SEO, INTELLIPAAT,Graduate from VIT University(IT),Vellore,Tamil Nadu, India.
Mr. Siddamreddy Sai Sreekar Reddy, VIT University(IT),Vellore, Tamil Nadu, India.

ABSTRACT :
The internet is the world's largest and most lavish data source. Search Engines are routinely used to retrieve information from the World Wide Web. Conventional search engines offer a straightforward interface for searching for user queries and providing results in the form of the web URL of the appropriate web page, but finding relevant information using traditional search engines has become quite difficult. This study presented a search engine that uses Machine Learning to provide more relevant web pages at the top of search results for user queries.
Keywords *:* Search Engine, Machine Learning.

I INTRODUCTION

The World Wide Web is a collection of distinct systems and servers linked together using various technologies and approaches. Every site has a large number of site pages that are created and delivered to the server. So, whenever a user requires anything, he or she must first input a term. A keyword is a group of words derived from a user's search query. A user's search input might be syntactically wrong. This is where the genuine necessity for search engines arises. Search engines offer a simple interface for searching and displaying user queries.
Crawler for the web Web crawlers assist in the collection of information about a website and the connections that lead to it. We exclusively use web crawlers to gather data and information from the internet and store it in our database.

Indexer An indexer is a programme that organises each phrase on each web page and saves the resulting list of words in a massive database.

It is mostly used to respond to a user's keyword and provide the most effective result for that search. The Page ranking algorithm in the query engine rates the URL using several algorithms in the query engine.

In this work, Machine Learning Techniques are used to get the best appropriate web URL for a given term. The PageRank algorithm's output is sent into the machine learning algorithm as input.

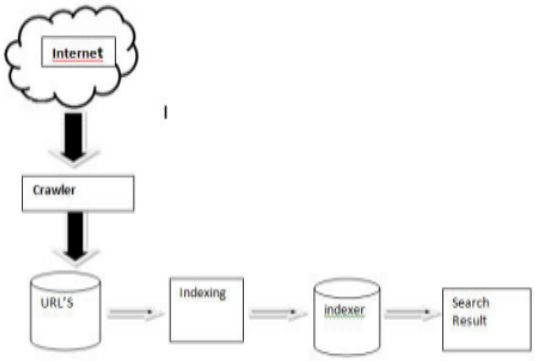


Fig1: System Diagram

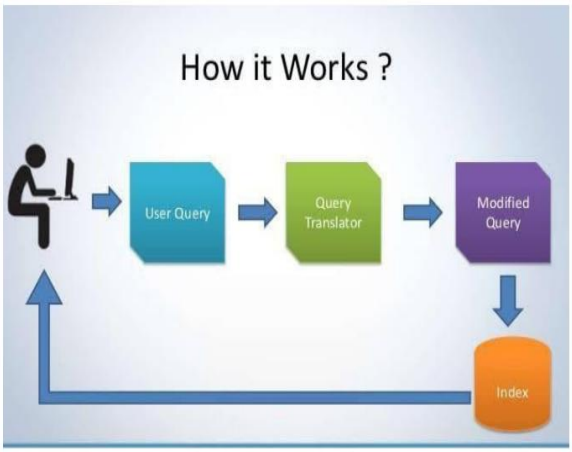


Fig 2: Working steps of search engine.

Three primary techniques are used by search engines: Web crawlers are bots that scour the internet for new web sites on a regular basis. Crawler gathers the information required to properly index sites and uses hyperlinks to go to other pages and index them as well.

A search index is a database of all online pages that is arranged in such a manner that keyword words and page content may be linked.

The quality of material in search engines' indexes may be classified in a variety of ways.

Algorithms for searching: They decide how search results are sorted based on quality and popularity by calculating the quality of web sites, determining how relevant a page is to a search phrase, and calculating how relevant a page is to a search word. To keep vast numbers of people coming back time after again, search engines strive to give the most beneficial results possible. This makes business sense, since most search engines rely on advertising revenue to stay afloat. In 2018, Google earned a staggering $116 billion.

OBJECTIVES AND GOALS

a) Goal

The goal of this research is to create a search engine that displays the URL of the most relevant website at the top of the search results, based on user queries. Our system's primary goal is to create a search engine that use machine learning techniques to improve accuracy over existing software.

b) The goal

The project's goal is to develop a search engine that displays the internet URL of the most relevant website at the top of the search results, based on user queries. Our system's primary goal is to create a search engine that use machine learning techniques to improve accuracy over existing software. Our system's primary goal is to create a search engine that uses machine learning techniques to improve accuracy over existing search engines**.**

PROBLEM STATEMENT

Correct and exact downfall forecast is still absent, which might help in a variety of industries like as agriculture, water conservation, and flood prediction. The problem is to come up with calculations for a downfall forecast that are backed by prior discoveries and similarities and provide output predictions that are both accurate and acceptable.

II. LITERATURE SURVEY

In the realm of search engines, data specialists and academics have made several efforts. Dutta and Bansal explore many types of search engines, concluding that the crawler-based search engine is the best of them all, and that it is also used by Google. It provides a user with a more appropriate web link in response to their inquiry. A web crawler is a software that navigates the internet by following a constantly changing, dense, and widely distributed hyperlinked structure, then saving downloaded pages in a large database that is then indexed for efficient execution of user queries.

The main advantage of utilising a keyword focused web crawler over a standard web crawler, according to the author [2, is that it operates intelligently and effectively. According to the user's needs, the search engine utilises a page ranking algorithm to place more relevant web pages at the top of the results. It simplifies the search process and allows the user to quickly get the information they seek. Initially, an idea was formed since users were having difficulty finding data, therefore a basic algorithm was devised that works on link structure. Later, as the web grew, weighted Page Rank and HITS were incorporated into the situation.

The author evaluates several Page Rank algorithms in [3,] and finds that the Weighted Page Rank method is the best fit for our system.

Hsinchun Chen and Michael Chau [4] describe a web page filtering system based on a machine learning technique. When the results of machine learning are compared to the results of conventional algorithms, it is discovered that the results of machine learning are more beneficial. The described method may potentially be used to create a search engine.

The author discusses a comparative study of page ranking algorithms in [5] and provides accurate results for user search phrases.

The author discusses a system for creating a domain specific search engine in [6]. Domain specific search engines are becoming more popular because they provide increased accuracy and extra functionality not available with general Web-wide search engines, such as multifaceted queries by age group, size, location, and cost. Machine learning methods are being used to automate the creation and upkeep of Domain Specific Search Engines.

III SYSTEM ANALYSIS

EXISTING SYSTEM :

Shodan is a PC software that allows you to search the internet for anything. Unlike Google and other search engines, Shodan indexes almost everything else besides the internet — internet cameras, water treatment facilities, yachts, medical devices, traffic lights, wind turbines, registration code readers, smart TVs, refrigerators, and anything else you can think of that's clogged into the internet (and sometimes shouldn't be). Open-port services, of course, use banners to advertise themselves. A banner communicates to the whole internet what service it provides and, as a result, the thanks to use it. Alternative services on other ports provide service-specific information, but there is no assurance that the printed banner is accurate or actual. In most circumstances, it is, and mercantilism, as a flag of purposeful dishonesty, is security via obscurity. Some businesses request that Shodan not locomotion their network, and Shodan complies. Attackers, on the other hand, don't require Shodan to find susceptible devices on your network. Shodan may cause temporary shame, but it is unlikely to improve your security posture.

PROPOSED SYSTEM

The software is taught using two supervised machine learning techniques, namely choice-based and review-based. To rank the links inside the coaching data-set, the tags/weights are computed. Each algorithm employs a variety of heuristics to get the same result. The frequency of the keyword in the link's content, as well as the location where it appears, determine the link's load. Heuristics such as whether the keyword is put in bold or italics; location wherever it appears, such as in the page title, headers, data, and so on; and the number of outbound links with the keyword in the address.

IV IMPLEMENTATION

SYSTEM ARCHITECTURE:

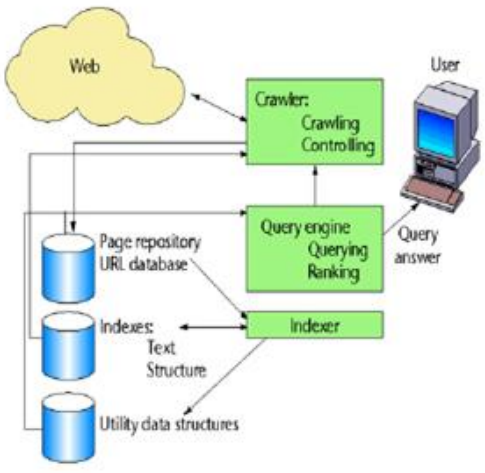


Fig.3: System Architecture

MODULES:

- Manager
- user
- Admin

- Machine-learning

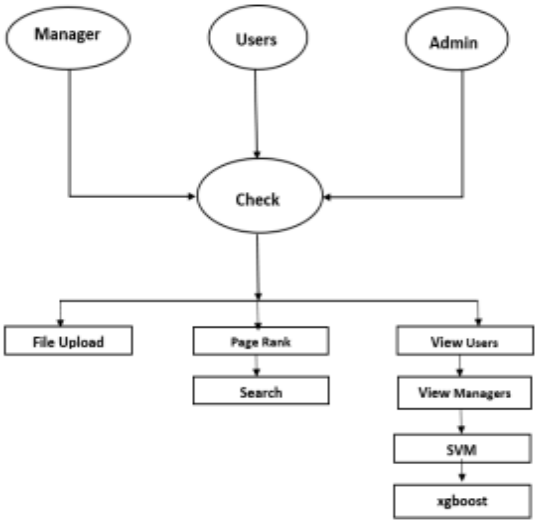


Fig 4: Modules

Manager:

For the whole experiment, manager information and job descriptions are provided. The file may be uploaded into the database by the manager. We may upload a file with the file type and name, as well as a specific url to the file, to get information about it.

User:

For the duration of the experiment, user information and task descriptions are provided. After logging into the session, the user will be presented with two alternatives. He can look up any website or piece of information he wants. Using the tf idf notion, we can search for a certain file and also determine its weight and rank.

Admin:

Managers and users will be given power by the administrator. In order to make it easier for managers and users to get started. All users and managers' information are visible to the administrator. The accuracy results of the svm and xgboost algorithms may be obtained by the administrator.

Machine learning:

Machine learning is a subfield of artificial intelligence (AI) that allows systems to learn and develop on their own without having to be explicitly programmed. Email spam and filtering, online fraud detection, and product recommendations are all examples of machine learning applications. Learning may be divided into three categories. The following are the details:

• supervised learning
• unsupervised learning
• reinforcement learning

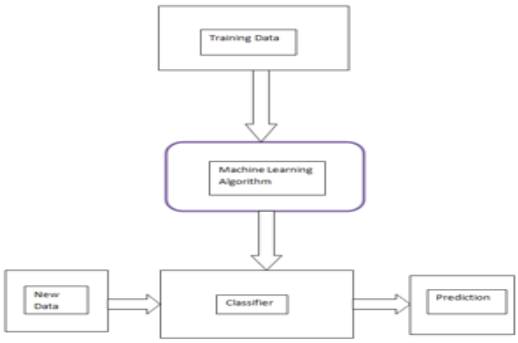


Fig 5: Machine Learning

Applications of Machine Learning

in Internet Search Engines Machine Learning may be used in a variety of ways when it comes to search engines. They are:

• Detecting patterns
• Recognizing new signals
• Image search to interpret photographs
• Custom signals depending on a particular question
• Improve ad quality by identifying similarities between terms in a search query • Query understanding
• URL/Document comprehension
• Search features
• Crawling
• User categorization
• Search Ranking
• Synonyms Identification/Query Expansion

Introduction to Search Engines

A search engine is a service that enables users to search for information on the internet (www). A search engine is a piece of software that looks for websites based on the words you provide as search criteria. They search their own databases of information to locate the information you're searching for. For search engines, there are three essential components. There are three of them: a web crawler, a database, and search interfaces.

Spiders and bots are other names for web crawlers. It's a piece of software that collects data from the internet. All of the material on the internet is saved in a database. It is made up of a large number of online resources. The search interface serves as a link between the user and the database. It assists users in searching the database.

Classification of Search Engines

Users may utilise a variety of search engines on the web depending on their use and functionality. Every search engine has a huge number of online pages in its database, however search engines with a high number of web pages are not the best. The top search engines will be those that provide accurate information based on the requested keyword. The following are the different types of search engines:

• Human-powered directories
• Meta search engines
• Hybrid search engines
• Specialty search engines
• Crawler-based search engines

Importance of Search Engine

Search engines are simply puzzles for the vast amount of data accessible on the internet. They enable consumers to swiftly and easily obtain material that is of true interest and value without having to wade through a plethora of irrelevant web sites. Users get search results from search engines that lead to relevant information on high-quality websites. Search engines are important because they are increasingly determining the information that customers access online regarding businesses, goods, and services. Being easy to locate on Google, Yahoo, and MSN is just as important as having a strong presence in print and broadcast media, as well as a successful conventional direct marketing campaign. To gain and maintain market share in online searches, search engines must establish a strong first impression by delivering results that are relevant to the user's search terms. They achieve this by archiving web pages that they create by collecting data using automated programmes known as "spiders" or "robots."

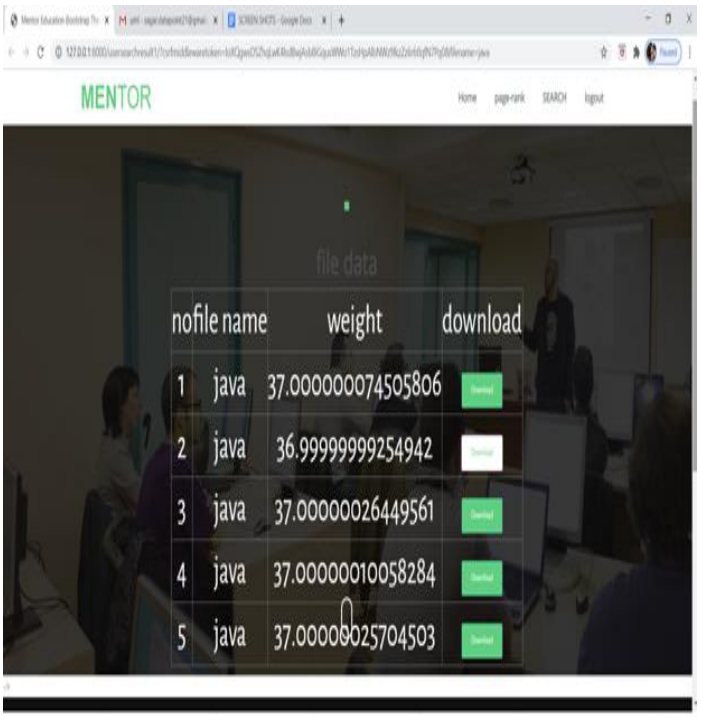
Search Engine Working

While you should always write content for your consumers rather than search engines, it's crucial to know how search engines function. Crawling is the method through which search engines like Google, Yahoo, and others locate new sites to index, and this is how most search engines develop their indexes. Bots or spiders are mechanisms that cruise the internet seeking for new sites. Typically, bots start with a list of websites. Previous crawls were used to identify URLs. They add these pages to the list of sites to index when they discover new links on these pages using tags like HREF and SRC. Then, depending on the search phrases you provided, search engines utilise their algorithms to generate a prioritised list from their index of which sites you should be most interested in. The engine will then produce a list of web results that have been rated according to its algorithm. Other variables on Google, such as customised and universal results, may affect your page ranking. The search engine uses extra information about the user to offer customised results that are tailored to their specific interests. Universal search results include video, photos, and Google news to produce a larger picture result, which might imply more competition for the same terms from other websites. Search engine optimization (SEO) is a set of guidelines that website owners may use to optimise their sites for search engines and hence enhance their rankings. The following are the steps in search engine optimization:

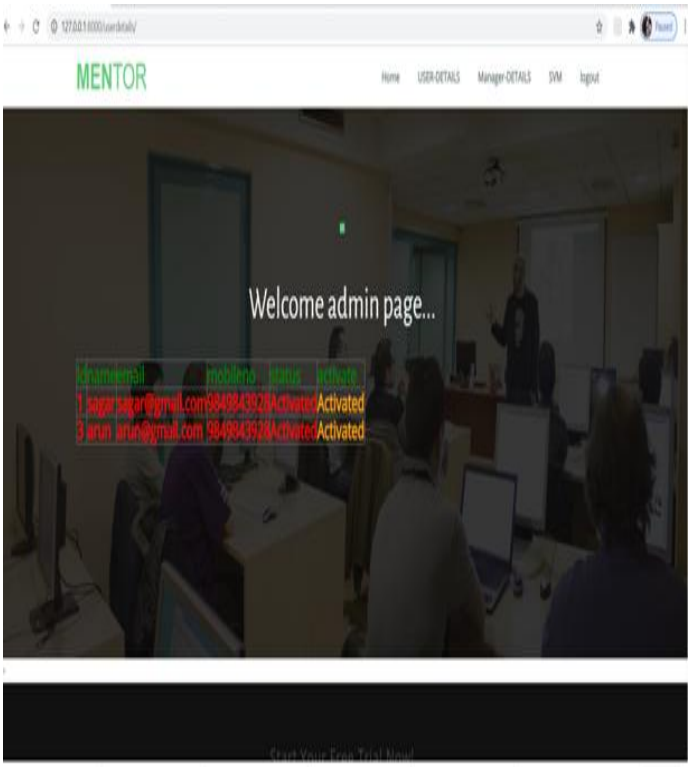
• Client needs
• Website analysis
• Keyword research
• Content creation
• Website optimization
• SEO submission
• Link building
• Reporting

V RESULT AND DISCUSSION

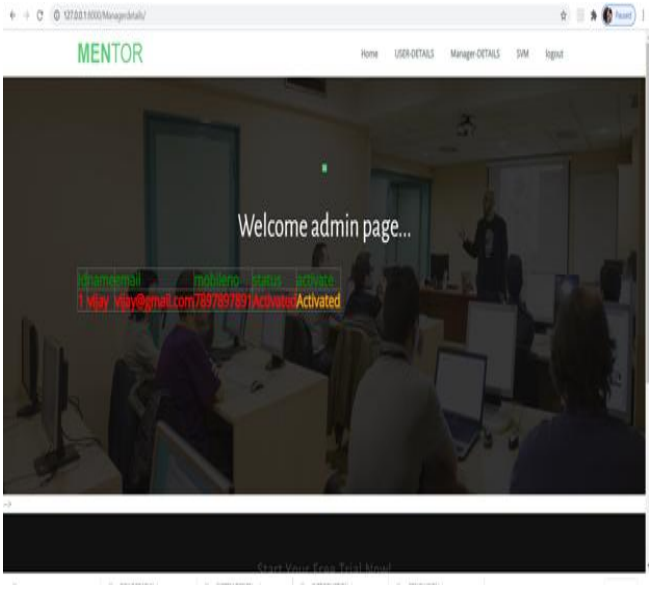
Login Papge:



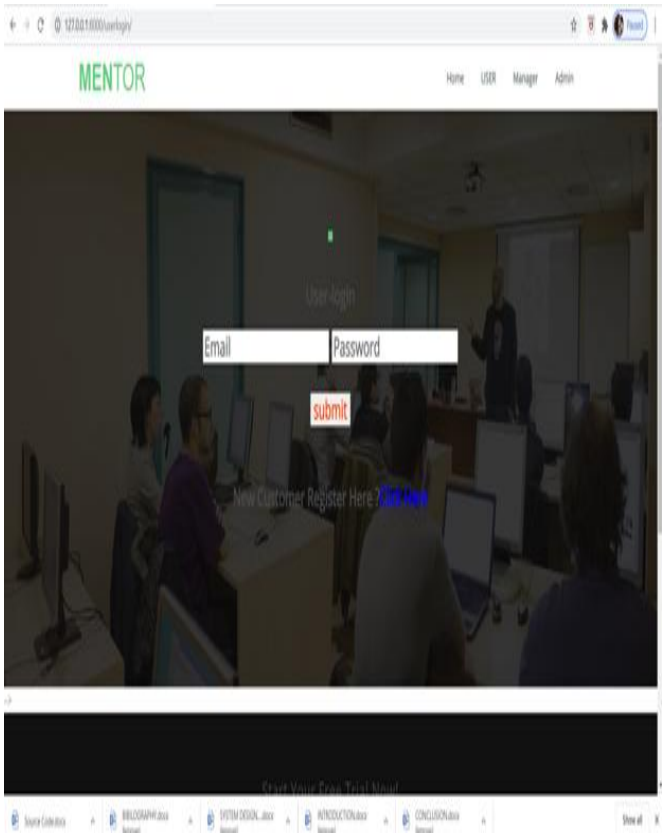
User details:



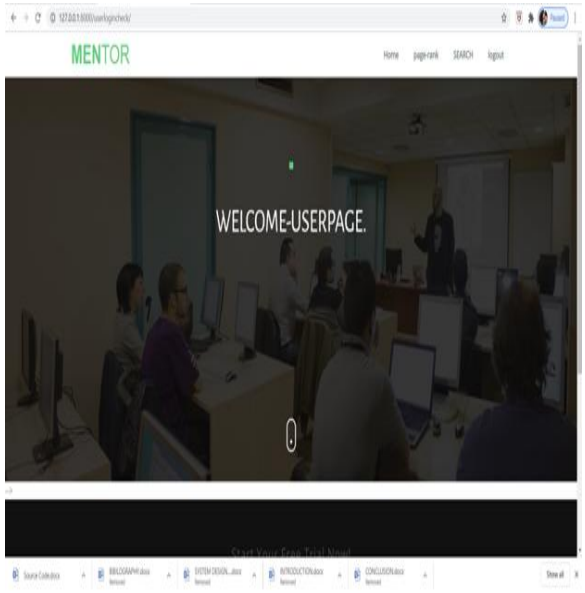
Manager-details:



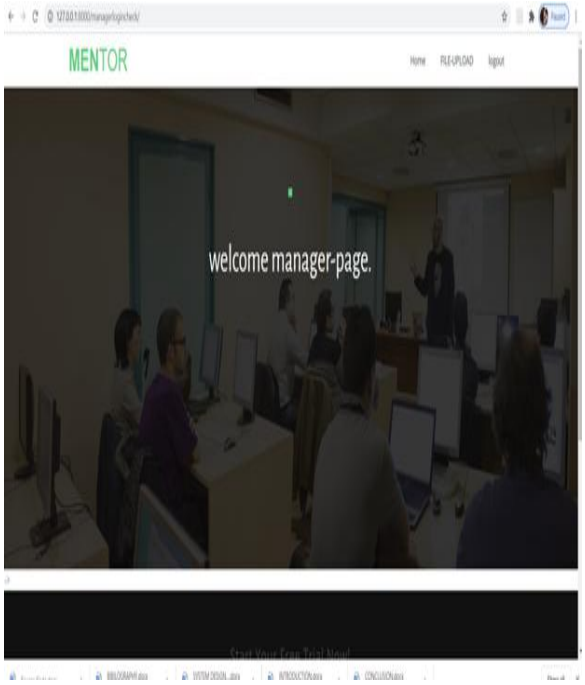
User login:



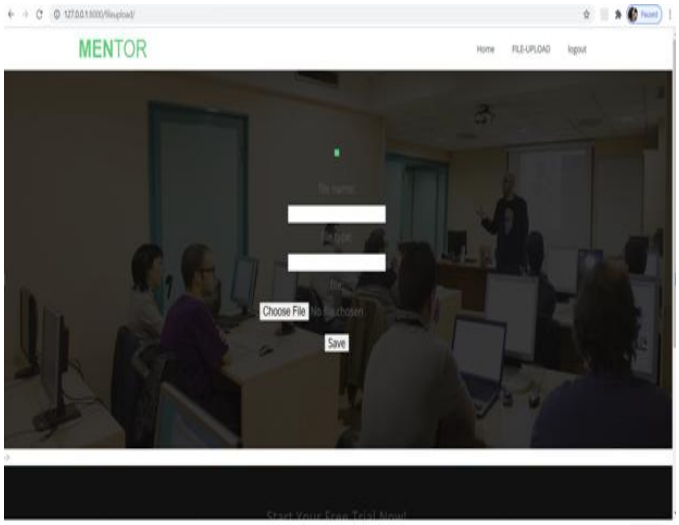
User home:



Manager home:



File upload:



VI CONCLUSION

For locating more relevant URLs for provided keywords, search engines are quite handy. As a result, the amount of time it takes for a user to find a suitable web page is lowered. Accuracy is a critical component in this. From the above, it can be inferred that XGBoost outperforms SVM and ANN in terms of accuracy. As a result, search engines based on the XGBoost and PageRank algorithms will be more accurate.

VII REFERENCES

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

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

[3] Tuhena Sen and Dev Kumar Chaudhary, "A Comparative Study of Simple, HITS, and Weighted PageRank Algorithms: Review," IEEE International Conference on Cloud Computing, Data Science & Engineering, 2017.

[4] Hsinchun Chen and Michael Chau, "A machine learning method to web page filtering utilising content and structural analysis," Decision Support Systems 44 (2008) 482–494, scienceDirect, 2008.

[5] Taruna Kumari, Ashlesha Gupta, and Ashutosh Dixit, "Comparative Study of Page Rank and Weighted Page Rank Algorithm," February 2014, International Journal of Innovative Research in Computer and Communication Engineering

[6] K. R. Srinath, "Page Ranking Algorithms - A Comparison," IRJET, Dec2017.

[7] S. Prabha, K. Duraiswamy, and J. Indhumathi (2014), "Comparative Analysis of Different Page Ranking Algorithms," International Journal of Computer and Information Engineering.

[8] A. K. Sharma and Dilip Kumar Sharma, "A Comparative Analysis of Web Page Ranking Algorithms," International Journal of Computer Science and Engineering, 2010.

[9] Vijay Chauhan, Arunima Jaiswal, and Junaid Khalid Khan, "Web Page Ranking Using Machine Learning Approach," 2015 International Conference on Advanced Computing and Communication Technologies.

[10] Tiewei s. Liu and Amanjot Kaur Sandhu, "Wikipedia Search Engine: Interactive Information Retrieval Interface Design," International Conference on Industrial and Information Systems, 2014.

[11] Neha Sharma, Rashi Agarwal, and Narendra Kohli, "Review of features and machine learning algorithms for online searching," 2016 International Conference on Advanced Computing and Communication Technologies.

[12] Sweah Liang Yong, Markus Hagenbuchner, and Ah Chung Tsoi, "Ranking Web Pages Using Machine Learning Approaches," 2008 International Conference on Web Intelligence and Intelligent Agent Technology

[13] "Weighted Page Rank Algorithm based on In-Out Weight of Webpages," Indian Journal of Science and Technology, Dec-2015. B. Jaganathan, Kalyani Desikan, "Weighted Page Rank Algorithm based on In-Out Weight of Webpages," Indian Journal of Science and Technology, Dec-2015.