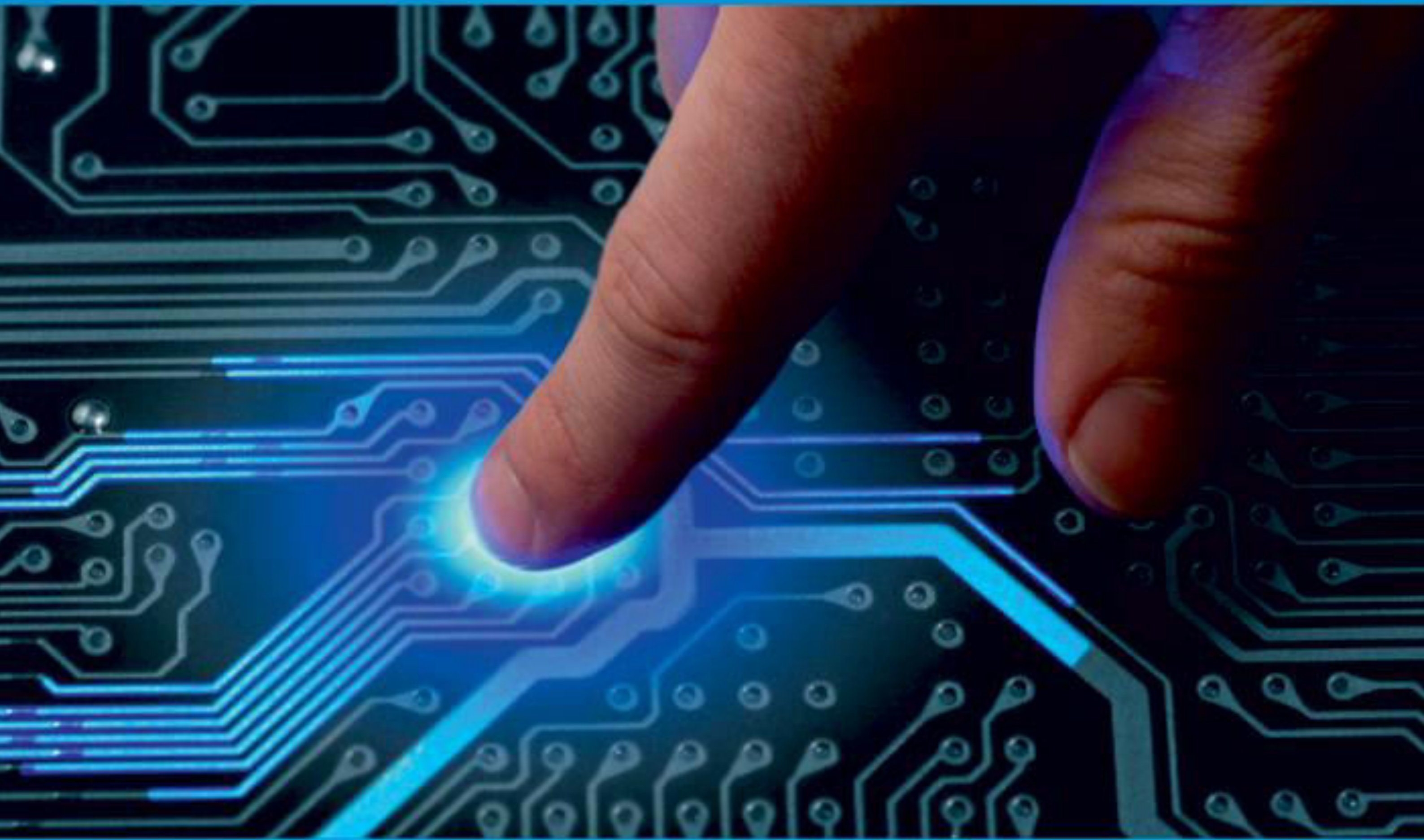
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 11, November 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 11, November 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1011015 |

Survey Paper on Various Techniques of Search Engine Using Machine Learning and Deep Learning

Aniket Sharma

UG Student, Dept. of B.Tech Computer Science, MIT World Peace University, Pune, India

ABSTRACT: The relevance of a web page is an innately biased matter and based on readers knowledge, interests and attitudes, web page is depended. There is still plenty to say that is accurate regarding the relative significance of web sites. The rapid expansion of the internet is one aspect that makes it challenging for search engines to provide consumers with relevant results within a set amount of time. Search engines employ classified directories to store the webpages, and some search engines even rely on human skill in this process. The majority of online pages classify web pages using automated techniques. Based on experimental findings, we may conclude that the best and most relevant strategy for search engines is to automatically classify web pages using machine learning techniques. The largest and most opulent source of information is the internet. To recover the information from World Wide Web, Search Engines are commonly utilized. Search engines offer a straightforward user interface for searching for user queries and providing results in the form of the web URL of the pertinent web page, but it has grown increasingly difficult to get the right information using conventional search engines. This study suggests a search engine that prioritizes more pertinent web sites for user searches and uses machine learning.

KEYWORDS: Search Engines, Expertise, Machine Learning, Deep Learning, Web Pages, Automated

I. INTRODUCTION

Search engines are employed to quickly and accurately find information on websites. Prior to the invention of search engines, finding necessary information on the internet was impossible. We may define a search engine as a piece of software that looks for websites using the words that users provide as search parameters. The field of the internet and search engines greatly benefits from search engine optimization. Google is the firm that was most successful in launching a search engine, as it made it easy and precise to search for online sites. Nowadays, the majority of search engines employ machine learning algorithms to automatically classify websites and rank websites. Different domains or topics that are connected to search engines and web page ranking can use machine learning.

The World Wide Web is basically a network of separate systems and servers that are linked together using various technologies and procedures. Numerous website pages are created and delivered to the server for each website. Therefore, a user must type a term if they need something. A keyword is a group of words that are taken from user-inputted search terms. User-provided search input might be syntactically wrong. This is where search engines actually become necessary. Search engines offer a straightforward user interface for searching user queries and display the results as the web address of the appropriate web page.

Three basic search engine components are highlighted in Figure 1.

1. Web crawler – It assist in gathering information about a website and the links that are connected to it. Web crawlers are the sole tools we use to collect data and information from the internet and store it in our database.
2. Each phrase on each web page is organized by an indexer, which then stores the resulting list of terms in a massive repository.
3. Search Engine It primarily serves to respond to user input and display the most useful results for that input.

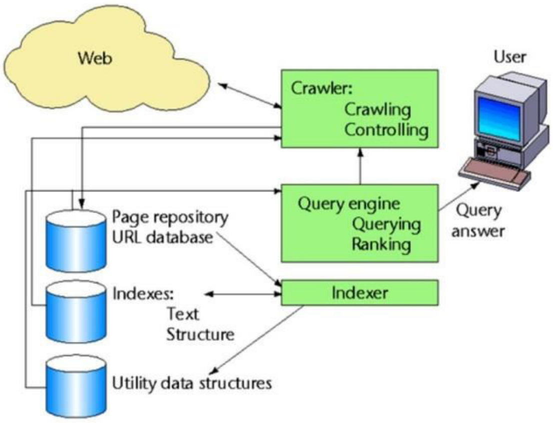


Fig 1. Block Diagram of Search Engine

This study makes use of machine learning techniques to identify the most appropriate website URL for the supplied keyword. Machine learning algorithms get the output of the PageRank algorithm as input. discusses related work being done using the PageRank algorithm and explains the goal. It discusses the machine learning-based suggested system, with section V containing the conclusion.

II. Related work

Data professionals and academics have made a lot of efforts in the search engine industry. After discussing numerous search engine types, Dutta and Bansal [8] come to the conclusion that the crawler-based search engine, which Google also utilizes, is the best one. It provides a user with a more pertinent site URL for their search. A Web crawler is a program that uses the frequently evolving, dense, and distributed hyperlinked structure to browse the internet. It then stores pages it has downloaded in a sizable database that is later indexed to effectively respond to user queries. The main advantage of using a keyword-focused web crawler over a standard one, according to the author in [9], is that it operates intelligently and effectively.

According to user demand, the search engine uses a page ranking algorithm to display more relevant web pages at the top of results. It makes searching easier and makes it simple for users to find the information they need. A simple method that relies on link structure was first introduced as a solution to a user problem with data searching, and as the web continued to grow, weighted PageRank and HITS were added to the equation. The Weighted PageRank algorithm is the greatest fit for our system out of all the PageRank algorithms that the author compares in [10]. For web page filtering, Michael Chau and Hsinchun Chen [11] suggested a solution that is based on a machine learning approach. When the outcomes of machine learning are compared to those of classical algorithms, it is discovered that the latter's outcomes are more beneficial. The suggested strategy works well for creating search engines as well.

A well-known method to rate the web pages accessible online is the webpage ranking algorithm. It helps us understand how the search engine exactly operates and how a machine learns itself while prioritizing the pages, determining which pages are crucial for successfully completing the user query need and which pages are less valuable. [12] The use of machine learning techniques also aids in our comprehension of the intricate page priority criteria used by the most well-known search engines, including Google, Yahoo, AltaVista, Dog pile, and numerous more. The web's structure was mostly revealed by page ranking.

The internet has become a vital source of knowledge for everyone due to its phenomenal growth in use [13]. Because of this, search engines play a crucial role in helping internet users find information. There has been a lot of research interest in the study of search engine user behavior. These studies are helpful in three ways: for users on a personal level, for government and marketing on a social society level, and for the development of more efficient search engines. These studies can be carried out by looking at the search engine's log files, which record user interactions with the search engine.

III. DETAILS OF TECHNOLOGY

A. *Search Engine*:

Internet users can utilize Search Engine [1] to conduct content searches on the World Wide Web (www). A search engine is a piece of software that looks for websites using the words you provide as search parameters. They search their own informational databases to see if they can find what you're looking for. For search engines, there are primarily three elements, they are:

- web crawler
- database
- search interfaces

Spiders and bots are other names for web crawlers. It is a piece of software that searches the internet for data. Every piece of information on the internet is kept in a database. There are tonne of online resources in it. An intermediary between the user and the database is the search interface. It makes it easier for people to search the database.

B. *Classification of Search Engines:*

Users can utilize any of the several search engines available on the web, depending on their capabilities and usage. Every search engine has a significant number of web pages kept in their database, but those search engines are not the top search engines. The best search engines will be those that will deliver accurate results for a given query. Search engines are classified as follows:

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

C. *Crawler Based Search Engines:*

Crawler-based search engines [3] like Google automatically build their listings. People search through what they have discovered once the web has been crawled or spidered. Changes to your webpages will be discovered by crawler-based search engines, which may have an impact on how you are ranked. Crawler-based search engines include three components:

- Crawler or spider
- index or catalog
- search engine software

The index or catalogue is like a big book that contains a copy of every webpage that the crawler or spider finds. The crawler or spider visits webpages and reads them. This book is updated whenever a webpage is updated.

D. *Meta Search Engines*:

The listings in a human powered directory like the open directory are provided by humans. With this kind of search engine, the website owner gives a brief description of the site together with the category it should be included in. The submitted site is then manually examined and either accepted for listing or refused for the appropriate category. The description of the sites will be compared to the keywords supplied in the search box. This indicates that modifications made to web page content are not taken into account because just the description is important. Compared to a site with weak content, a good site with good material is more likely to receive a free review.

| DOI: 10.15680/IJIRCCE.2022.1011015 |

E. *Search Engine Working*:

It's crucial to comprehend how search engines operate even though you should always write content for your website that is focused on your clients rather than search engines. Crawling, the method used by search engines like Google, Yahoo, and others to identify new pages to index, is the foundation upon which the majority of them build their indexes. Bots or spiders are devices that search the web for new pages. Usually, the bots begin with a list of websites. URLs identified by earlier crawls. They add theses to the list of sites to index when they discover new links on these pages thanks to tags like HREF and SRC. Then, based on the search terms you entered; the search engine utilizes its algorithms to produce a prioritized list from its index of the pages you should be most interested in. The engine will then deliver a list of online results that have been ranked according to its unique methodology. Other factors on Google, such as personalized and universal results, could also affect how your page ranks. In personalized results, the search engine makes use of additional user data to deliver results that are specifically tailored to the user's preferences. Greater competition from other websites for the same terms can result from universal search results, which mix video, photos, and Google news to produce a bigger picture result.

A set of guidelines called search engine optimization can be used by website owners to better optimise their websites for search engines and raise their search engine rankings. Additionally, it's a terrific approach to improve the usability, speed, and navigation of your website, which will raise its quality. The following are the steps in search engine optimization:

- Website analysis
- Client requirements
- Keyword research
- Content writing
- Website optimization
- SEO submission
- Link building
- Reporting

F. *Applications of Machine Learning in Search Engines*:

There are several search engine-related applications for machine learning. Those are:

- Pattern Detection
- Identifying new signals
- Custom signals based on specific query
- Image search to understand photos
- Identifying similarities between words in a search query
- Improve ad quality
- Query understanding
- URL/Document understanding
- Search features [2]
- Crawling [1]
- User classification
- Search Ranking
- Synonyms Identification/Query Expansion
- Intent Disambiguation

G. *Experimental Results*:

The list of implemented algorithms is provided below. The PageRank algorithm utilises the algorithm that provides greater accuracy.

I. Support Vector Machine
II. Artificial Neural Network

III. XGBoost

Support Vector Machine

A SVM was also employed to enable a better method due to its excellent performance. It classified data using the same set of feature scores. We are using nonlinear SVM because the dataset cannot be separated linearly. Nonlinear kernels come in the forms of Rbf, Poly, and Sigmoid. The aforementioned 14 features were chosen as inputs for the SVM model, which then attempted to predict whether or not each web page in the testing set was relevant to the supplied query based on those features. The outcomes were saved and applied to performance assessment.

Artificial Neural Network

The input layer, hidden layer, and output layer are the three layers that make up a neural network. The input layer of the neural network had 14 nodes, one for each of the 14 feature scores for each web page. The output layer just needs one output node to determine whether a web page is relevant. In the buried layer, there were set to be 7 nodes. A grid search is used to set these values after some preliminary testing. The batch size is set to 10 and the entire process has been run 150 times. The outcomes were saved and applied to performance assessment.

XGBoost

It is a form of ensemble learning based on boosting. In order to increase precision and speed, it employs gradient-boosted decision trees. We use a gbtree-based booster and the input feature has the same 14 features. 50 classifiers are in use, and the maximum depth size is set at 4. Based on some preliminary experimentation, these parameters are set utilising a parameter turning and cross validation method.

IV. Conclusion

Using a search engine to identify more relevant URLs for a particular keyword is incredibly helpful. As a result, it takes less time for users to find the appropriate web page. Accuracy is a key component in this. The aforementioned observation leads to the conclusion that XGBoost outperforms SVM and ANN in terms of accuracy. The accuracy of a search engine created with the XGBoost and PageRank algorithms will therefore be higher.

References

1. Shams, A.B.; Hoque Apu, E.; Rahman, A.; Sarker Raihan, M.M.; Siddika, N.; Preo, R.B.; Hussein, M.R.; Mostari, S.; Kabir, R. Web Search Engine Misinformation Notifier Extension (SEMiNExt): A Machine Learning Based Approach during COVID-19 Pandemic. Healthcare 2021, 9, 156.
2. R. Karwa and V. Honmane, "Building Search Engine Using Machine Learning Technique," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019
3. Matošević, G.; Dobša, J.; Mladenić, D. Using Machine Learning for Web Page Classification in Search Engine Optimization. Future Internet 2021, 13, 9.
4. McCallum, Andrew, et al. "A machine learning approach to building domain-specific search engines." IJCAI. Vol. 99. 1999.
5. Tuarob, Suppawong, Prasenjit Mitra, and C. Lee Giles. "Building a search engine for algorithms." ACM SIGWEB Newsletter (2014)
6. Neenu Ann Sunny,Machine Learning in Search Engines,ISSN: 2321-9939 | ©IJEDR 2020 Year 2020, Volume 8, Issue 2
7. Boyan, Justin, Dayne Freitag, and Thorsten Joachims. "A machine learning architecture for optimizing web search engines." AAAI Workshop on Internet Based Information Systems. 1996.
8. 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.
9. MACHINE LEARNING BASED SEARCH ENGINE WITH CRAWLING, INDEXING AND RANKING, DOI: 10.47760/ijcsmc.2021.v10i07.011

10. 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.
11. Michael Chau, Hsinchun Chen, "A machine learning approach to web page filtering using content and structure analysis", Decision Support Systems 44 (2008)
12. Vishwas Ravall and Padam Kumar, "SEReleC (Search Engine Result Refinement and Classification) – A meta search engine based on combinatorial search and search keywordbased link classification," in IEEE-International Conference on advances in Engineering, science and management (ICAESM-2012), March 30,31,2012. Saad ALBAWI, Tareq Abed MOHAMMED, Saad AL-ZAWI, "Understanding of a convolutional neural network," ICET 2017.
13. FarzanehShoeleh, Mohammad Sadegh Zahedi, MoiganFarhoodi, "Search Engine Pictures: Empirical analysis of a web search engine query log," in Third International Conference on Web Research(ICWR), 19,20 April 2017.

BIOGRAPHY

Aniket Ranjeet Sharmais aundergraduate student in the Computer Science Engineering Department, MIT World Peace University, Pune. His research interests are Blockchain technology, Machine Learning, Algorithms, Cyber Security, etc.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.165

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details