HBRP PUBLICATION

Journal of Advance Research in Mobile Computing
Volume 3 Issue 2

Using Machine Learning To Build A Search Engine

Priyanka R[1], Megha G S[2*]
[1]Student, [2]Assistant Professor
Department of Computer Science Engineering,
East West College of Engineering, Bangalore, Karnataka, India.
****Corresponding Author***
E-mail Id:-*meghags@ewce.edu.in*

ABSTRACT

The Internet is a massive server and the most preferred abundant data source. We use search engine as a popular method to retrieve information from the internet. A search engine is a website through which users can search the content of the Internet. It is one of the primary ways that internet users find to obtain suitable information. Now a days search engine providers grows in popularity because they offer increased accuracy and extra functionality which is not possible in the general. Searching for information on the internet differs in several ways. In this paper we propose Page Ranking (PR), Weighted PR(WPR) and Hyperlink Induced Topic Search (HITS) algorithms using machine learning technique to greatly automate the methods and classification of Web pages. Search engines play a critical role in the growth of the internet; they assist many internet users in quickly finding relevant information. It can be used to do the basic process of retrieving information.

Keywords:-*Search engine, page ranking, weighted page rank, hits, machine learning.*

INTRODUCTION

Machine learning is the study of computer algorithms that improve themselves over time as a result of their experiences. It's an Artificial Intelligence subfield. Machine Learning is a modern innovation that has assisted man in improving not only various industrial and professional procedures, but also everyday life.. Machine learning contributed their work in many fields like Speech Recognition, Image Recognition, Medical Diagnosis, Learning Association, Prediction System, Financial Services etc.

Because there are 8 billion web pages available, searching for information that is especially needed would be difficult. Search engines are used to filter the information on the internet and translate it into results in a couple of seconds .Finding requirement information on web was unfeasible before search engine were introduced. A search engine is a piece of software that searches the web for terms you specify as search terms. Because each search engine has its own catalogue or data base of various types of information, you will obtain different /hits if you use multiple search engines. The deep web is a term used to describe internet content that cannot be searched using a web search engine. WWB is a network of independent systems and servers that are linked together using various technologies and approaches. Search engine is an automated software program and a dedicated website to search other website and contents .

GOOGLE, YAHOO, ASK.COM, and other search engine tools are powered by search engine software that allows the database to be searched. A search engine is a software programme that allows users to type in keywords and obtain information from websites. It only searches when a user asks a search engine for information,

not the entire internet. It's possible that one search engine will yield results/hits that another won't.

Figure 1: focuses on 3 main components of a search engine

Crawler or Spider

Information is actually collected from different pages of the website available on the web

Indexer

Is a huge server is actually a place where the collected space crawled information is actually stored.

Parser or Retrieval

When the whole content is stored after indexing, the information will be picked up and displayed for the user.

This research employs machine learning techniques to find the almost perfect web URL for a given key word. The output of the page rank algorithm is fed into the machine learning algorithm as an input.

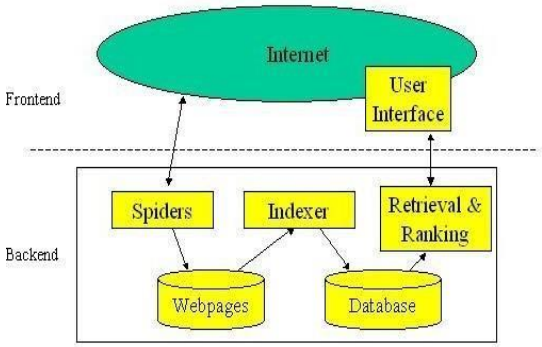

Fig.1:- Components of a Search Engine

LITERATURE SURVEY

Manika Dutta and K.L. Bansal [1] Discusses various types of search engines and come to the conclusion that the crawler-based search engine, which Google also uses, is the best of them all. Presents a person with additional site addresses that are relevant to their search. A web crawler is a programme that navigates the internet by saving downloaded pages while following the constantly changing dense and widely spread hyperlinked structure.

Gunjan H, Nikita V.Mahajan [2] According to the author, the most significant advantage of using a keyword focused web crawler over a regular web crawler is that it performs wisely and effectively. A page ranking algorithm is used by the search engine to provide results. A more relevant web page, according to Google, is at the top of the search results. The user's requirements It streamlines the search experience and guarantees that the user gets the information they're looking for. More changes were made to expand Weighted PageRank, and HITS entered the picture.

Tuhena Sen, Dev Kumar [3] After comparing various PageRank algorithms, the author concludes that the Weighted PageRank method is the best fit for our system.

Michael Chau, Hsinchunn Chen [4] A web page filtering system based on machine learning was presented. When the findings of machine learning were compared to those of a traditional algorithm, it was determined that machine learning provided
better outcomes. are more advantageous. The suggested technique can also be used to build asearch engine.

OBJECTIVE

Create a search engine that, based on user queries, presents the web URL of the most relevant web page at the top of search results. Our system’s major purpose is to construct a search engine that improves accuracy over existing search engines by utilizing machine learning techniques.

METHODOLOGY

The technique for developing a search engine is outlined below in step-by-step format.

Using a web crawler to collect data fromthe internet:

We acquire data and information from the

internet using a keyword-based web crawler in this step. A crawler starts with a list of URLs to visit, then follows every hyperlink it finds on each page and adds it to the list. Web data crawlers are primarily employed to make a copy of all the pages visited so that they can be processed later in a search engine.

Clean up your data:

*Data cle*aning is done in this step to pre-process the data and remove any extraneous information. After gathering data from the internet using a web crawler, data cleaning is required, which includes tokenization, capitalization, removing stop words, identifying parts of speech, and lemmatization.

Comparing the existing algorithms:

PageRank (PR):

PageRank is a metric for determining how important a website's pages are. Google search uses it to rank webpages in their search engine results.

iiWeighted Page Rank(WPR):

Instead of splitting the rank value of a page, it assigns higher rank values to more important pages.

Hyperlink Induced Topic Search(HITS):

Is a website ranking system based on link analysis. This algorithm is applied to the structure of web links. The Weighted PageRank algorithm is the greatest fit for the system since it provides better accuracy and efficiency than other algorithms.

In Machine Learning, combine the chosen method with the best feature:

After selecting and implementing the most appropriate PageRank algorithm for our needs. In this phase, the machine learning algorithm uses the topmost result of the PageRank algorithm, and the output of the machine learning algorithm is sent to the user as a web address of a relevant web page based on the user's request.

Implementing and displaying an efficient user query result:

Finally, create a query engine that takes the user's feedback in the form of a query and displays the most effective results for that query. Based on the performance of the machine learning algorithm, it will show the web address of related websites.

Output:

Following is a list of the algorithms that have been introduced. With the PageRank algorithm, the algorithm that provides more accuracy is used.

Support Vector Machine: It is usually used in classification as a supervised machine learning technique. Problems that make it possible to take a better approach to its expected performance.

Artificial Neural Network: Are a sort of machine learning algorithm that is modelled after the human brain, and works similarly to how neurons in our nervous system can learn from past data and respond with predictions or classifications.

XGBoost:

eXtreme Gradiant Boosting (XGBoost) is a gradient boosted decision tree solution aimed for speed and performance.

ACCURACY OF DIFFERENT ALGORITHM

$$accuracy = \frac{number\ of\ documents\ correctly\ classified}{total\ number\ of\ documents}$$

No.	Algorithm	Accuracy
1	SVM	89.50
2	ANN	91.35
3	XGBoost	92.59

CONCLUSION

For finding more appropriate url's for a given keyword, a search engine is extremely useful. As a result, user time spent searching for relevant web pages is reduced, and accuracy is a very important

factor. Based on the above observations, it can be concluded that XGBoost is more accurate than SVM and ANN. Search engine built using XGBoost and PageRank algorithm will provide betteraccuracy.

REFERENCES

1. Karwa, R., & Honmane, V. (2019, May). Building search engine using machine learning technique. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (pp. 1061-1064). IEEE.
2. Liang, C. (2011, July). User profile for personalized web search. In *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* (Vol. 3, pp. 1847-1850). IEEE.
3. Hongqing, G., Peiyong, S., Wenzhong, G., & Kun, G. (2018, November). Component-based Assembling Tool and Runtime Engine for the Machine Learning Process. In *2018 International Conference on Cloud Computing, Big Data and Blockchain (ICCBB)* (pp. 1-7). IEEE..
4. Oyama, S., Kokubo, T., & Ishida, T. (2004). Domain-specific web search with keyword spices. *IEEE Transactions on knowledge and data engineering*, *16*(1), 17-27.
5. Usta, A., Altingovde, I. S., Ozcan, R., & Ulusoy, O. (2021). Learning to Rank for Educational Search Engines. *IEEE Transactions on Learning Technologies*.
6. Alchalabi, A. E., Elsharnoby, M., & Khawaldeh, S. (2016, July). Rightscope: Detecting search campaingns positive and negative queries. In *2016 International Conference on Machine Learning and Cybernetics (ICMLC) 1*. 290-295. IEEE.
7. Lu, H., Su, S., Tian, Z., & Zhu, C. (2019). A novel search engine for Internet of Everything based on dynamic prediction. *China Communications*, *16*(3), 42-52.
8. Wang, Z., Wang, Q., & Wang, D. (2006, October). Application of domain-specific search method in meta-search engine on internet. In *The Proceedings of the Multiconference on" Computational Engineering in Systems Applications" 2.* 2078-2085. IEEE.
9. Hatcher, W. G., Qian, C., Gao, W., Liang, F., Hua, K., & Yu, W. (2021). Towards Efficient and Intelligent Internet of Things Search Engine. *IEEE Access*, *9*, 15778-15795.
10. Liang, F., Qian, C., Hatcher, W. G., & Yu, W. (2019). Search engine for the internet of things: Lessons from web search, vision, and opportunities. *IEEE Access*, *7*, 104673-104691..
11. Yan, J., Zhao, Z. X., Xu, N. Y., Jin, X., Zhang, L. T., & Hsu, F. H. (2012, April). Efficient query processing for web search engine with FPGAs. In *2012 IEEE 20th International Symposium on Field-Programmable Custom Computing Machines* (pp. 97-100). IEEE.
12. Ding, Z., Gao, X., Guo, L., & Yang, Q. (2012, November). A hybrid search engine framework for the internet of things based on spatial-temporal, value-based, and keyword-based conditions. In *2012 IEEE International Conference on Green Computing and Communications* (pp. 17-25). IEEE.
13. Menon, K. D., Raj Jain, A., & Kumar Pareek, D. (2019). Quantitative analysis of student data mining.
14. Pai H, A., HS, S., Soman, S., Pareek, D., & Kumar, P. (2019). Analysis of causes and effects of longer lead time in software process using FMEA.
15. Pai H, A., HS, S., Soman, S., Pareek, D., & Kumar, P. (2019). ROC Structure Analysis of Lean Software Development in SME's Using

Mathematical CHAID Model.

16. HS, S., Soman, S., & Kumar Pareek, D. (2019). Fast and efficient parallel alignment model for aligning both long and short sentences.
17. BR, M., Bhavya, B. R., Pareek, D., & Kumar, P. (2016). Education Data Mining: Perspectives of Engineering Students. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN*, 2347-5552
18. Kotagi, M., & Pareek, P. K. (2016). Survey on Challenges in DevOps. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN*, 2347-5552.
19. Soman, S., & Pareek, P. K. (2020). An exploratory analysis on challenges prevailing in small and medium IT firms. In *Journal of Physics: Conference Series* (Vol. 1427, No. 1, p. 012010). IOP Publishing.
20. Sangeetha, V., Vaneeta, M., Kumar, S. S., Pareek, P. K., & Dixit, S. (2021). Efficient Intrusion detection of malicious node using Bayesian Hybrid Detection in MANET. In *IOP Conference Series: Materials Science and Engineering* .1022, No. 1, p. 012077). IOP Publishing.Pareek, P.
21. K., Swathi, K., & Shetteppanavar, P. (2017, May). An efficient machine translation model for Dravidian language. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* . 2101-2105. IEEE.
22. Aditya Pai, H., Pareek, P. K., Narasimha Murthy, M. S., Dixit, S., & Karamadi, S. (2021). An Exploratory Study for Process Optimization in IT Industry. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, 3* . 617-631. Springer Singapore.
23. Soman, S., Pareek, P. K., Dixit, S., Chethana, R. M., & Kotagi, V. (2021). Exploration Study to Study the Relationships Between Variables of Secure Development Lifecycle (SDL). In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, Volume 3* (pp. 641-649). Springer Singapore.
24. Suhas, G. K., Devananda, S. N., Jagadeesh, R., Pareek, P. K., & Dixit, S. (2021). Recommendation-Based Interactivity Through Cross Platform Using Big Data. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, 3* (pp. 651-659). Springer Singapore
25. Soman, S., Pareek, P. K., Dixit, S., Kotagi, V.. (2020). An Empirical Investigation on Practicing Secure Software Development in Software Development Life Cycle in Small & Medium Level Software Firms in Bengaluru. *International Journal of Advanced Science and Technology, 29*(7s), 5164.
26. Patil, S. S., Pareek, P.,K., Dinesh, H. A., Arlimatti, S.(2017). Review of relay selection techniques in multi-hop wireless sensor network with iot. *International Journal of Creative Research Thoughts (IJCRT), 5*(4).846-850