BUILDING SEARCH ENGINE USING MACHINE LEARNING TECHNIQUES

[1]Mr. K. Venkateswara Rao, [2]S. HARISH REDDY, [3]R. ANVESH, [4]S. VANDANA

[1]Associate Professor, Department of Information Technology, CMR College of Engineering & Technology

[2, 3, 4] B-Tech, Department of Information Technology, CMR College of Engineering & Technology

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. We proposed a search engine using Machine Learning technique that will give more relevant web pages at top for user queries.

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.

Problem Statement

The world's population is widely increasing day by day. Almost every person uses the internet and smart technology. Because of the widespread use of web pages nowadays, retrieving information from the internet presents a significant challenge. The complexity of getting the results is increasing. Maintaining and understanding the data becomes very complex. The accuracy of results is low due to a lack of algorithms. It provides 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. So, to overcome this problem, we are building a search engine using machine learning.

OBJECTIVE:

Numerous endeavors have been made by data experts and researchers in the field of search engine. Dutta and Bansal [1] discuss

various type of search engine and they conclude the crawler based search engine is best among them and also Google uses it. A Web crawler is a program that navigates the web by following the regularly changing, thick and circulated hyperlinked structure and from there on putting away downloaded pages in a vast database which is after indexed for productive execution of user queries. In [2], author conclude that major benefit of using keyword focused web crawler over traditional web crawler is that it works intelligently, efficiently. The search engine uses a page ranking algorithm to give more relevant web page at the top of result, according to user need.Initially just an idea has been developed as user were facing problem in searching data so simple algorithm introduced which works on link structure, then further modification came as the web is also expanding so weighted PageRank and HITS came into the scenario. In [3], author compare various PageRank algorithm and among all, Weighted PageRank algorithm is best suited for our system. Michael Chau and Hsinchun Chen [4] proposed a system which is based on a machine learning approach for web page filtering. The machine learning result is compared with traditional algorithm and found that machine learning result are more useful. The proposed approach is also effective for building a search engine.

Project Scope and Limitations

The primary limitation of this work is that the keywords and associated webpages came from only one company and industry. In addition, the chosen language (Finnish) might affect the results. In general, the gift industry can be considered as a highly competitive online industry with a lot of SEO activity taking place. Although the range of keywords was relatively large in the context of that company, as was the number of webpages, this research would need to be replicated using data on other companies, industries, and languages in order to claim generalizability of the findings. Even though we mitigate the impact of potential personalization by using an anonymous browser, there are other factors that impact the searchresults, such as click logs, ranking information from past SERPs, and so on. These factors make search results structurally unstable and make it more difficult to replicate research in this domain. Moreover, as the ranking algorithms of the major search engines undergo periodic changes, any research in the SEO field is subject to expiration. Even with the mentioned limitations, the results are indicative of the impact of content and link features on search rankings. Acquiring

more data would allow for the use of more features (e.g., utilizing unsupervised methods such as topic modeling), and more learning examples to further improve the algorithm. In addition, more features about the actual content of the sites, would provide more distinct information about each site. Apart from obtaining data from other contexts, future research could focus on specific website elements. In particular, the relatively high correlation of H3 and rankings is an interesting finding. One reason for this can be that the use of H3 tags is rarer than the use of H1 and H2 tags and, therefore, websites using H3 tags are applying more advanced SEO and content marketing strategies. This proposition should be explored infuture research.

IMPLEMENTATION

Smart Track utilizes GS1 standards barcodes containing unique serialized product identifier, Lot production and expiration dates. The information contained in the GS1 barcode is captured across various supply chain processes and used to maintain a continuous log of ownership transfers. As each stakeholder records the possession of the product, an end user (patient) can verify authenticity through central data repository maintained as Global Data Synchronization Network (GDSN) by using a smartphone app. In the downstream supply chain at the warehouse, pharmacy and hospital units can scan the barcode to verify the product and its characteristics.

Introduction

BERT is an open source machine learning framework for natural language processing (NLP). BERT is designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context. The BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with question and answer datasets.BERT, which stands for Bidirectional Encoder Representations from Transformers, is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection. (In NLP, this process is called *attention*.)Historically, language models could only read text input sequentially -- either left-to-right or right-to-left -- but couldn't do both at the same time. BERT is different because it is designed to read in both directions at once. This capability, enabled by the introduction of Transformers, is known as bidirectionality. Using this bidirectional capability, BERT is pre-trained on two different, but related, NLP tasks: Masked Language Modeling and

Next Sentence Prediction.The objective of Masked Language Model (MLM) training is to hide a word in a sentence and then have the program predict what word has been hidden (masked) based on the hidden word's context. The objective of Next Sentence Prediction training is to have the program predict whether two given sentences have a logical, sequential connection or whether their relationship is simply random.

Merits, Demerits and Challenges

- **Merits**
- Much better model performance over legacy methods
- An ability to process larger amounts of text and language
- **Demerits**
- The model is large because of the training structure and corpus.
- It is slow to train because it is big and there are a lot of weights to update.
- It is expensive.

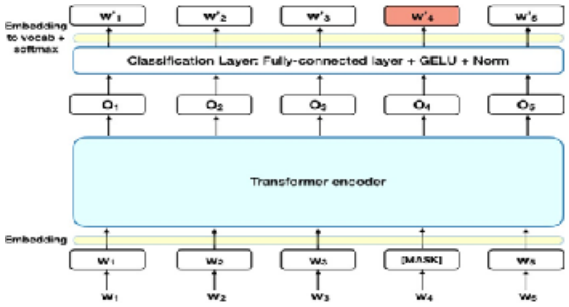

Figure 1: Structure of Search Engine using BERT

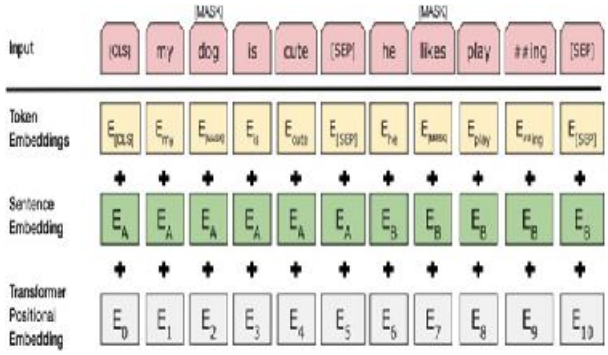

Figure 2: BERT with Modifications

PROPOSED SYSTEM

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

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.

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.

Working of a SVM algorithm:

One reasonable choice as the best hyperplane is the one that represents the largest separation or margin between the two class

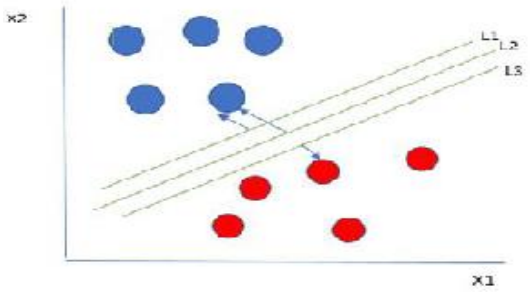

Fig:-3

So we choose the hyperplane whose distance from it to the nearest data point on each side is maximized. If such a hyperplane exists it is known as the maximum-margin hyperplane/hard margin. So from the above figure, we choose L2. Let's consider a scenario like shown below

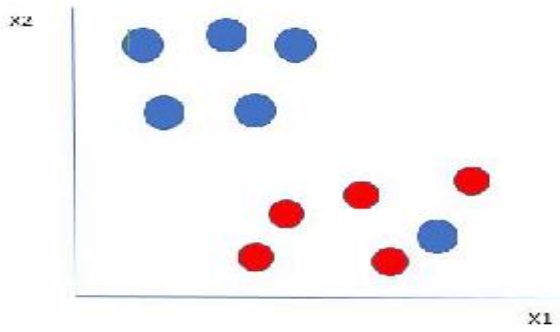

Fig:-4

RESULTS :

In this paper author is using machine learning algorithms called SVM and XGBOOST to predict search result of given query and building search engine with machine learning algorithms. To train this algorithm author is using website data and then this data will be converted to numeric vector called TFIDF (term frequency inverse document frequency). TFIDF vector contains average frequency of each words.

In this paper, author has implemented following modules

- **Admin module:** admin can login to application using username and password as admin and then accept

or activate new users registration and then train SVM and XGBOOST algorithm

- **Manager module:** manager can login to application by using username and password as Manager and Manager and then upload dataset to application
- **New User Signup:** using this module new user can signup with the application
- **User Login:** user can login to application and then perform search by giving query.

To run project install MYSQL and python 3.7 and then copy content from DB.txt file and paste in MYSQL to create database.

Now double click on 'run.bat' file to start python DJANGO server and get below screen

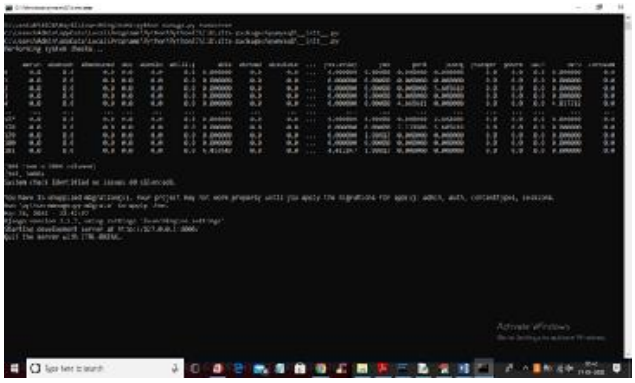

Fig:-5

In above screen server started and build a vector from dataset where first row showing word and remaining rows contains TFIDF word frequency. Now open browser and enter URL as http://127.0.0.1:8000/index.html and press enter key to get below page

SCREENSHOTS:

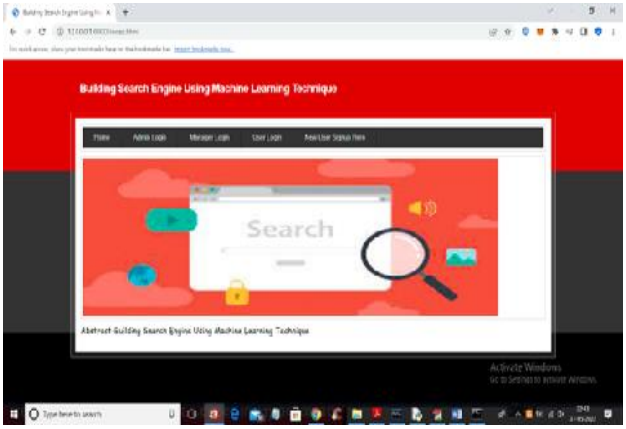

Fig:-6

In above screen click on 'New User Signup Here' link to get below screen

CONCLUSION

Search engine is very useful for finding out more relevant URL for given keyword. Due to this, user time is reduced for searching the relevant web page. Due to privacy reasons and other reasons we want to build own search engine. The project we have built is used to providethe faster retrieval of information using search engines that are implemented by using machine learning algorithms. It provides 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.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 Computingand 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] Revathy, G. & Gurumoorthi Elangovan, Dr & Sasikala, C. & Latha, T.. (2023). Training superbot with learning automata and multi kernel SVM. AIP Conference Proceedings. 020034. 10.1063/5.0154176.

[12] Prakash K, L.N.C., Surya Narayana, G., Ansari, M.D., Gunjan, V.K., 2022, Optimization of K-Means Clustering with Modified Spiral Phenomena, Lecture Notes in Electrical Engineering, 10.1007/978-981-16-7985-8_126

[13] Kumar, S., Chandra Sekhar Redd, L., George Joseph, S., Kumar Sharma, V., H, S., 2022, Deep learning based model for classification of COVID âˆ'19 images for healthcare research progress, Materials Today: Proceedings, 10.1016/j.matpr.2022.04.884

[14] Ramana, M.V., Rao, G.K.M., Reddy, M.D., Kumar, B.V.R.R., Kumar, P.R., 2022, Optimisation and impact of process parameters on tool-chip interaction while turning of A286 iron based nickel superalloy, International Journal of Machining and Machinability of Materials, 10.1504/IJMMM.2022.122783

[15] Das, S., Nayak, S.C., Sahoo, B., 2022, Modeling and Forecasting Stock Closing Prices with Hybrid Functional Link Artificial Neural Network, Smart Innovation, Systems and Technologies, 10.1007/978-981-16-9447-9_19

[16] Patel, P., Sivaiah, B., Patel, R., 2022, Relevance of Frequent Pattern (FP)-Growth-Based Association Rules on Liver Diseases, Lecture Notes in Networks and Systems, 10.1007/978-981-19-0901-6_58