INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 07, July 2023, pp : 267-276

e-ISSN : 2583-1062

Impact Factor : 5.725

BUILDING SEARCH ENGINE USING MECHINE LEARNING TECHNIQUE - REVIEW

P. Tejaswini[1], Saragadam Sridhar[2]

[1]Department Of Master Of Computer Science Miracle Educational Society Group Of Institutions Vizianagram– 535216 (AP) India

DOI: https://www.doi.org/10.58257/IJPREMS31799

ABSTRACT

Building Search Engine using Machine Learning Technique The web is the huge and most extravagant well spring of data. To recover the information from the WWW(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. This paper proposed a search engine using Machine Learning technique that will give more relevant web pages at top for user queries.In this paper we used XGBOOST algorithm.

Keywords— WWW(World Wide Web), XGBOOST (Extreme Gradient Boosting)

1. INTRODUCTION

World Wide Web is actuallya 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. • 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. • Indexer which arranges each term on each web page and stores the subsequent list of terms in a tremendous repository. • Query Engine 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. • This paper utilizes Machine Learning Techniques to discover the utmost suitable web address for the given keyword. The output of the Page Rank algorithm is given as input to the machine learning algorithm.

2. LITERATURE SURVEY

1)S. Su, Y. Sun, X. Gao, J. Qiu* and Z. Tian*. A Correlation-change based Feature Selection Method for IoT Equipment Anomaly Detection. Applied Sciences.

In the era of the fourth industrial revolution, there is a growing trend to deploy sensors on industrial equipment, and analyze the industrial equipment's running status according to the sensor data. Thanks to the rapid development of IoT technologies [1], sensor data could be easily fetched from industrial equipment, and analyzed to produce further value for industrial control at the edge of the network or at data centers. Due to the considerable development of deep learning in recent years, a common practice of such analysis is to conduct deep learning [2,3,4]. Such methods select a subset of all fetched sensor data stream as the input features, and generate equipment predictions. As a result, the performance of the learning model was seriously impacted by the features selected, thus feature selection plays a critical role for such methods.

2)To select an appropriate set of features for the learning model, researchers aim to select the most relevant features to the prediction model to improve the prediction performance, or to select the most informative features to conduct data reduction. Unfortunately, both kinds of methods have intrinsic drawbacks when applied in the online scenarios. The former kind of methods seriously depends on predefined evaluation criteria, such as feature relevance metrics [5] or a predefined learning model [6]. Thus, such method are limited to certain dataset, and are not suitable for online scenarios which involve dynamical and unsupervised feature selection. The later kind of methods right fits in the online scenarios. However, data reduction mainly aims to improve the efficiency (but not accuracy) of the prediction model, which is not the most concerning factor of online industrial equipment status analysis.

To relieve the dependency of predefined evaluation criteria, researchers switch to select the features which can indicate the online sensor data's characters, such as features which are smoothest on the graph [7], or the features with highest clusterability [8,9]. In this paper, we focus on the features with correlation changes such as smoothness and clusterability, which are important characters for traditional pattern recognition fields like image processing and voice

recognition [7,8,9]. We believe that correlation changes can significantly pinpoint status changes in industrial environment. As far as we know, this is the first work focusing on correlation changes for online feature selection.

3)X. Yu, Z. Tian, J. Qiu, F. Jiang. A Data Leakage Prevention Method Based on the Reduction of Confidential and Context Terms for Smart Mobile Devices. Wireless Communications and Mobile Computing,https://doi.org/10.1155/2018/5823439.

With the development of Internet and information technology, smart mobile devices appear in our daily lives, and the problem of information leakage on smart mobile devices will follow which has become more and more serious [1, 2]. All kinds of private or sensitive information, such as intellectual property and financial data, might be distributed to unauthorized entity intentionally or accidentally. And that it is impossible to prevent from spreading once the confidential information has leaked.

4)According to survey reports [3, 4], most of the threats to information security are caused by internal data leakage. These internal threats consist of approximate 29% private or sensitive accidental data leakage, approximate 16% theft of intellectual property, and approximate 15% other thefts including customer information, and financial data. Further, the consensus of approximate 67% organizations shows that the damage caused from internal threats is more serious than those form outside.

Although laws and regulations have been passed to punish various behaviors of intentional data leakage, it is still hard to prevent data leakage effectively. Confidential data can be easily disguised by rephrasing confidential contents or embedding confidential contents in nonconfidential contents [5, 6]. In order to avoid the problems arising from data leakage, lots of software and hardware solutions have been developed which are discussed in the following chapter.

In this paper, we present CBDLP, a data leakage prevention model based on confidential terms and their context terms, which can detect the rephrased confidential contents effectively. In CBDLP, a graph structure with confidential terms and their context involved is adopted to represent documents of the same class, and then the confidentiality score of the document to be detected is calculated to justify whether confidential contents is involved or not. Based on the attribute reduction method from rough set theory, we further propose a pruning method. According to the importance of the confidential terms and their context, the graph structure of each cluster is updated after pruning. The motivation of the paper is to develop a solution which can prevent intentional or accidental data leakage from insider effectively. As mixed-confidential documents are very common, it is very important to accurately detect the documents containing confidential contents even when most of the confidential contents have been rephrased.

5)Y. Sun, M. Li, S. Su, Z. Tian, W. Shi, M. Han. Secure Data Sharing Framework via Hierarchical Greedy Embedding in Darknets. ACM/Springer Mobile Networks an

Geometric routing, which combines greedy embedding and greedy forwarding, is a promising approach for efficient data sharing in darknets. However, the security of data sharing using geometric routing in darknets is still an issue that has not been fully studied. In this paper, we propose a Secure Data Sharing framework (SeDS) for uture darknets via hierarchical greedy embedding. SeDS adopts a hierarchical topology and uses a set of secure nodes to protect the whole topology. To support geometric routing in the hierarchical topology, a two-level bit-string prefix embedding approach (Prefix-T) is first proposed, and then a greedy forwarding strategy and a data mapping approach are combined with Prefix-T for data sharing. SeDS guarantees that the publication or request of a data item can always pass through the corresponding secure node, such that security strategies can be performed. The experimental results show that SeDS provides scalable and efficient end-to-end communication and data sharing.

6) Z. Wang, C. Liu, J. Qiu, Z. Tian, C., Y. Dong, S. Su Automatically Traceback RDP-based Targeted Ransomware Attacks. Wireless Communications and Mobile Computing. 2018.https://doi.org/10.1155/2018/7943586.

With the popularization of new energy electric vehicles (EVs), the recommendation algorithm is widely used in the relatively new field of charge piles. At the same time, the construction of charging infrastructure is facing increasing demand and more severe challenges. With the ubiquity of Internet of vehicles (IoVs), inter-vehicle communication can share information about the charging experience and traffic condition to help achieving better charging recommendation and higher energy efficiency. The recommendation of charging piles is of great value. However, the existing methods related to such recommendation consider inadequate reference factors and most of them are generalized for all users, rather than personalized for specific populations

3. PROPOSED METHOD

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

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 07, July 2023, pp : 267-276

e-ISSN : 2583-1062

Impact Factor : 5.725

frequency). TFIDF vector contains average frequency of each words.The proposed search engine is very useful for finding out more relevant URLs for given keywords

4. ADVANTAGES OF PROPOSED SYSTEM

- We will build a search engine which gives the 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 to discover the utmost suitable web address for the given keyword by using machine learning techniques for increasing accuracy compared to available search engines.

In this paper we are 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 we implemented the following modules

1)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

2)Manager module: manager can login to application by using username and password as Manager and Manager and then upload dataset to application

3)New User Signup: using this module new user can signup with the application

4)User Login: user can login to application and then perform search by giving query.

XG Boost

XG BOOST is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now. Please see the chart below for the evolution of tree-based algorithms over the years.

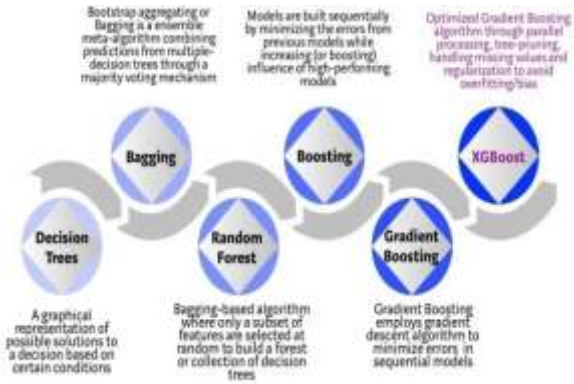

Fig-1

XG Boost algorithm was developed as a research project at the University of Washington. Tianqi Chen and Carlos Guestrin presented their paper at SIGKDD Conference in 2016 and caught the Machine Learning world by fire. Since its introduction, this algorithm has not only been credited with winning numerous Kaggle competitions but also for being the driving force under the hood for several cutting-edge industry applications. As a result, there is a strong community of data scientists contributing to the XGBoostopen source projects with ~350 contributors and ~3,600 commits on GitHub. The algorithm differentiates itself in the following ways:

1. A wide range of applications: Can be used to solve regression, classification, ranking, and user-defined prediction problems.
2. Portability: Runs smoothly on Windows, Linux, and OS X.
3. Languages: Supports all major programming languages including C++, Python, R, Java, Scala, and Julia.
4. Cloud Integration: Supports AWS, Azure, and Yarn clusters and works well with Flink, Spark, and other ecosystems.

XGBoost and Gradient Boosting Machines (GBMs) are both ensemble tree methods that apply the principle of boosting weak learners (CARTs generally) using the gradient descent architecture. However, XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 07, July 2023, pp : 267-276

e-ISSN : 2583-1062

Impact Factor : 5.725

Fig-2

Algorithmic Enhancements:

1. **Regularization**: It penalizes more complex models through both LASSO (L1) and Ridge (L2) regularization to prevent overfitting.
2. **Sparsity Awareness**: XG Boost naturally admits sparse features for inputs by automatically 'learning' best missing value depending on training loss and handles different types of sparsity patterns in the data more efficiently.
3. **Weighted Quantile Sketch:** XGBoost employs the distributed weighted Quantile Sketch algorithm to effectively find the optimal split points among weighted datasets.
4. **Cross-validation**: The algorithm comes with built-in cross-validation method at each iteration, taking away the need to explicitly program this search and to specify the exact number of boosting iterations required in a single run.

1) Advantages of XGBoost:

1. Performance: XG Boost has a strong track record of producing high-quality results in various machine learning tasks, especially in Kaggle competitions, where it has been a popular choice for winning solutions.
2. Scalability: XGBoost is designed for efficient and scalable training of machine learning models, making it suitable for large datasets.
3. Customizability: XGBoost has a wide range of hyperparameters that can be adjusted to optimize performance, making it highly customizable.
4. Handling of Missing Values: XGBoost has built-in support for handling missing values, making it easy to work with real-world data that often has missing values.

2) Disadvantages of XGBoost:

1. Computational Complexity: XGBoost can be computationally intensive, especially when training large models, making it less suitable for resource-constrained systems.
2. Overfitting: XGBoost can be prone to overfitting, especially when trained on small datasets or when too many trees are used in the model.
3. Hyperparameter Tuning: XGBoost has many hyperparameters that can be adjusted, making it important to properly tune the parameters to optimize performance. However, finding the optimal set of parameters can be time-consuming and requires expertise.
4. Memory Requirements: XG Boost can be memory-intensive, especially when working with large datasets, making it less suitable for systems with limited memory resources.

Fig-3

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

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 07, July 2023, pp : 267-276

e-ISSN : 2583-1062

Impact Factor : 5.725

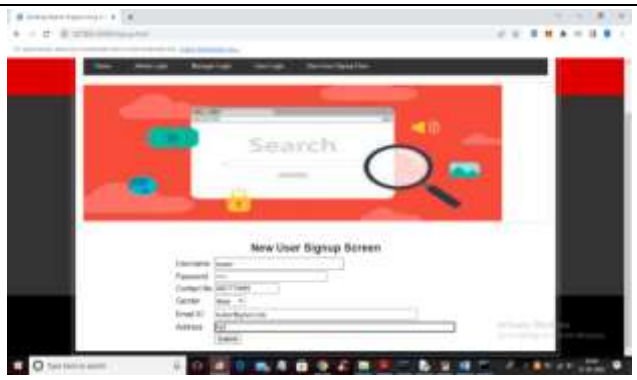

Fig-4

In above screen user is signing up and then press button to get below output

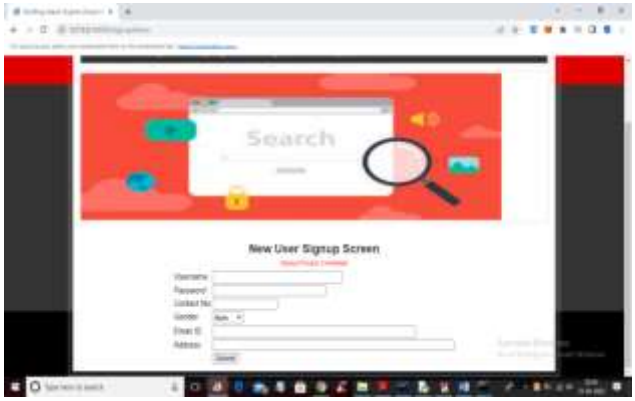

Fig-5

In above screen user signup process completed and now click on 'User Login' to get below screen

Fig-6

In above screen user is login and will get below output

Fig-7

In above screen we gave correct login but account not activated by admin so click on 'Admin Login' link to login as admin and then activate user

Fig-8

In above screen admin can click on 'View Users' link to view all users

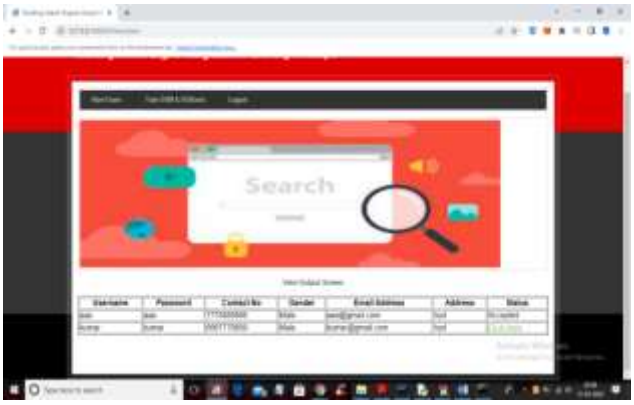

Fig-9

In above screen admin can click on 'Click Here' link to activate that user account

Fig-10

In above screen we can see admin activated kumar user account and now admin can click on 'Train SVM & XGBOOST' link to train machine learning SVM and XGBOOST algorithm and get below output.

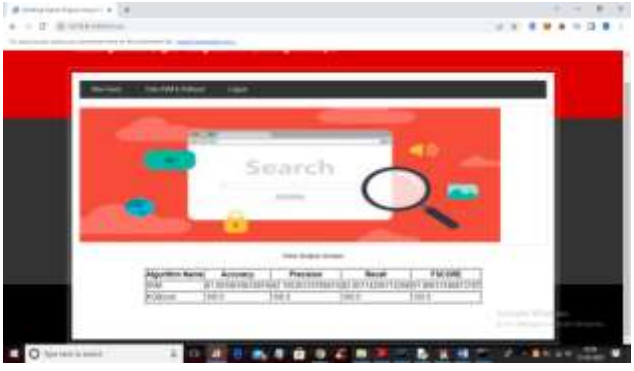

Fig-11

In above screen we can see SVM and XGBOOST accuracy and in both algorithms XGBOOST got high accuracy and now logout and login as Manager

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 07, July 2023, pp : 267-276

e-ISSN : 2583-1062

Impact Factor : 5.725

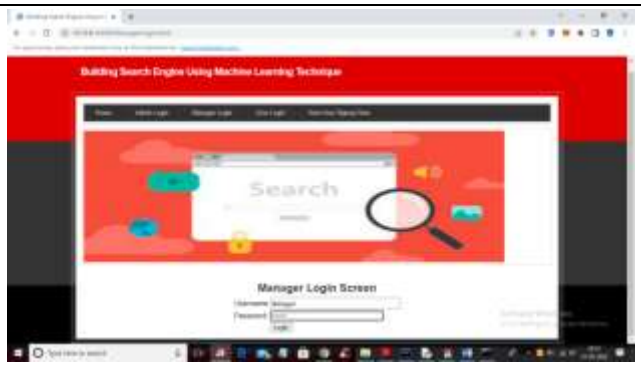

Fig-12

In above screen manager is login and after login will get below screen

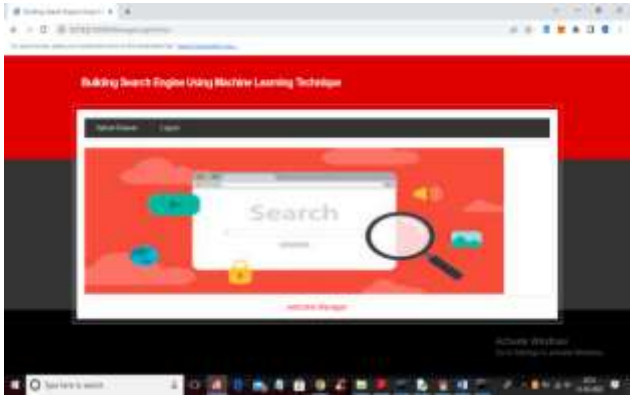

Fig-13

In above screen manager can click on 'Upload Dataset' link to upload dataset or documents

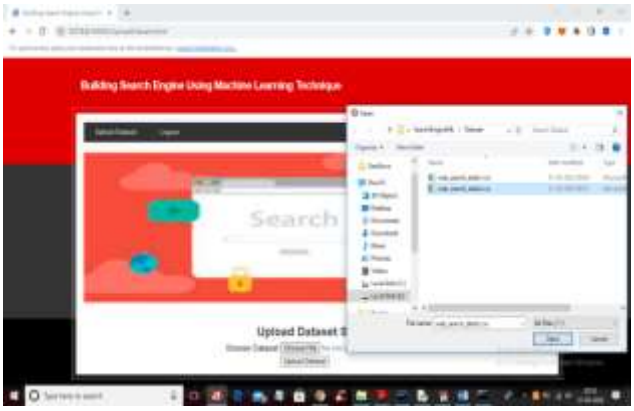

Fig-14

In above screen manager is browsing and uploading dataset and this file you can find inside 'Dataset' folder and now press button to saved dataset at server database

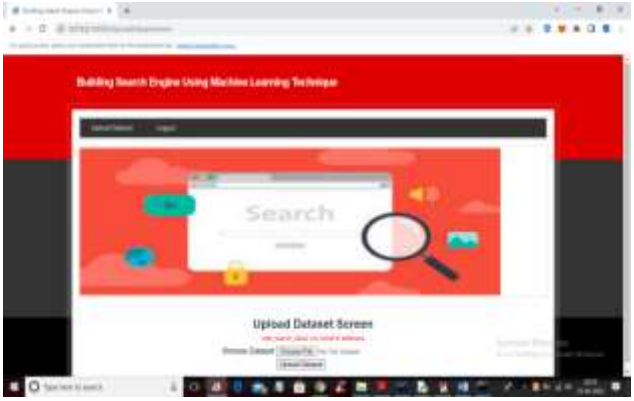

Fig-15

In above screen dataset file saved in database and now logout and login as user to perform search

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 07, July 2023, pp : 267-276

e-ISSN : 2583-1062

Impact Factor : 5.725

Fig-16

In above screen user is login and after login will get below output

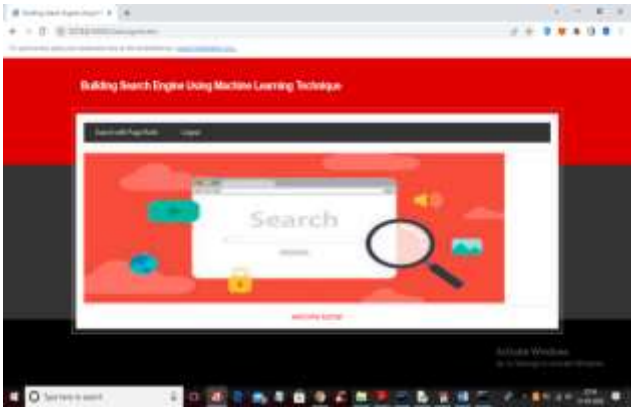

Fig-17

In above screen user can click on 'Search with Page Rank' link to search any data

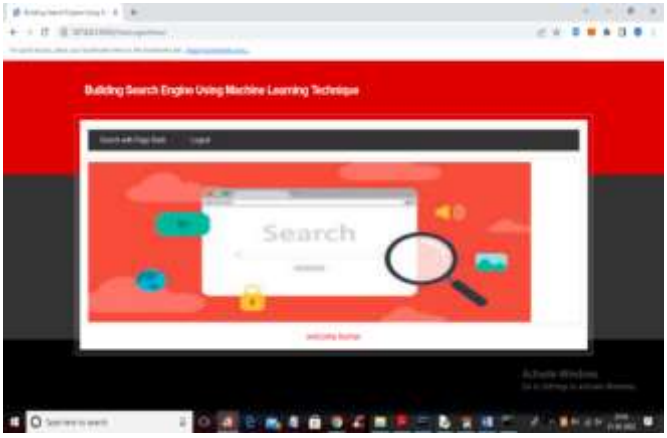

Fig-18

In above screen user can click on 'Search with Page Rank' link to search any data

Fig-19

In above screen I entered query as 'news on security' and press button to get below search result

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 07, July 2023, pp : 267-276

e-ISSN : 2583-1062

Impact Factor : 5.725

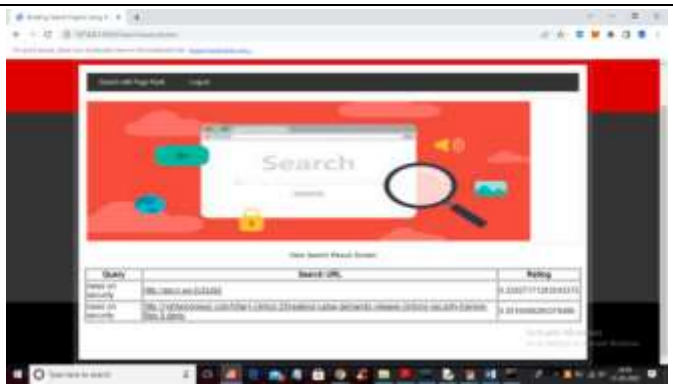

Fig-20

In above screen machine learning algorithm predicts two URLS for given query and user can click on those URLS to visit page

Fig-21

In above screen by clicking on URL link user can visit and view page. Similarly user can give any query and if query available in dataset then he will get output

Fig-22

For above query we got below result

5. CONCLUSION

search engines are very useful for finding out more relevant urls for given keywords. due to this, user time is reduced for searching the relevant web page. 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.

6. 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.

[14] Programming Python, Mark Lutz

[15] Head First Python,Paul Barry

[16] Core Python Programming, R. Nageswara Rao

[17] Learning With Python, Allen B. Downey

[18] https://www.w3schools.com/python/

[19] https://www.tutorialspoint.com/python/index.htm

[20] https://www.javatpoint.com/python-tutorial

[21] https://www.learnpython.org/

[22] https://www.pythontutorial.net/