

DEVELOPING A SEARCH ENGINE WITH MACHINE LEARNING TECHNIQUES

1.Dr. R. Konda Reddy, 2.G. Geetha Madhuri, 3.Y. Hima Chandana, 4.N. Manoj, 5.SK. Ameer Basha

#1 Professor in Department of CSE, PBR Visvodaya Institute Of Technology & Science (Autonomous), Kavali.

#2#3#4#5 B. Tech with Specialization of Computer Science and Engineering in PBR Visvodaya Institute Of Technology & Science (Autonomous), Kavali.

ABSTRACT_ Building a Search Engine using Machine Learning Techniques The internet is the largest and most extravagant source of data. Search engines are widely used to retrieve information from the Internet. Search engines provide a simple interface for searching for user queries and providing results in the form of the web address of the relevant web page, but utilizing traditional search engines to find relevant information has become quite difficult. This study proposed a search engine that uses Machine Learning techniques to display more relevant web sites at the top of user queries.

1.INTRODUCTION

Internet is really a trap of individual frameworks and servers which are associated with various innovation and strategies. Each site includes the stacks of site pages that are being made and sent on the server. So on the off chance that a client needs something, the individual in question requirements to type a watchword. Watchword is a bunch of words removed from client search input. Search input given by a client might be linguistically erroneous. Here comes the real requirement for web crawlers. Web crawlers give you a basic point of interaction to look through client questions and show the outcomes. • Web crawlers help in gathering information about a site and the connections connected with them. We are just involving web crawlers for gathering information and data from WWW and putting away it in our data set. • Indexer which orchestrates each term on each page and stores the ensuing rundown of terms in a colossal vault. • Question Motor is basically used to answer to the client's watchword and show the compelling result for their catchphrase. In the question motor, the Page positioning calculation positions the URL by involving various calculations in the inquiry motor. • This paper uses AI Procedures to find the highest level of reasonable web address for the given catchphrase. The result of the Page Rank calculation is given as contribution to the AI calculation.

2.LITERATURE SURVEY

2.1 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.

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.

2.2.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.

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.

2.3 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 and Security. 2018. <https://doi.org/10.1145/3211111>

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 future 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.

2.4 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

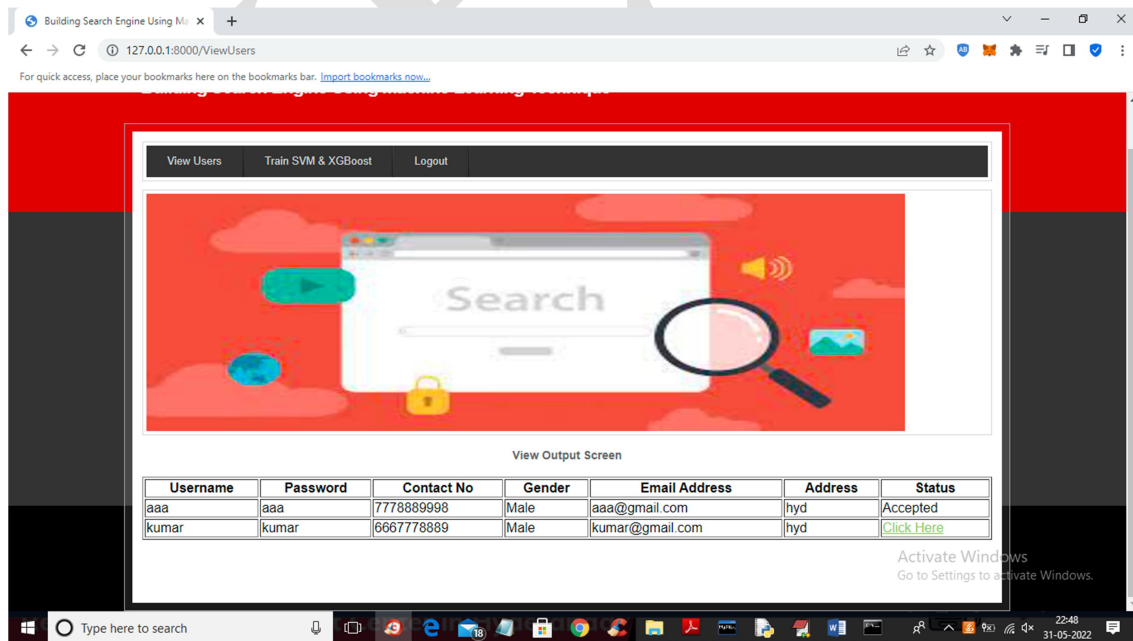
3.PROPOSED SYSTEM

In this research, the author uses machine learning methods known as SVM and XGBOOST to forecast search results for a given query and to design a search engine using machine learning algorithms. To train this method, the author uses website data, which is then turned into a numeric vector known as TFIDF (term frequency inverse document frequency). TFIDF vectors contain the average frequency of each word. The proposed search engine is highly helpful in locating more relevant URLs for provided keywords.

3.1 IMPLEMENTATION

- 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 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.

4.RESULTS AND DISCUSSION



Building Search Engine Using M... x

127.0.0.1:8000/ViewUsers

For quick access, place your bookmarks here on the bookmarks bar. [Import bookmarks now...](#)

View Users Train SVM & XGBoost Logout

Search

View Output Screen

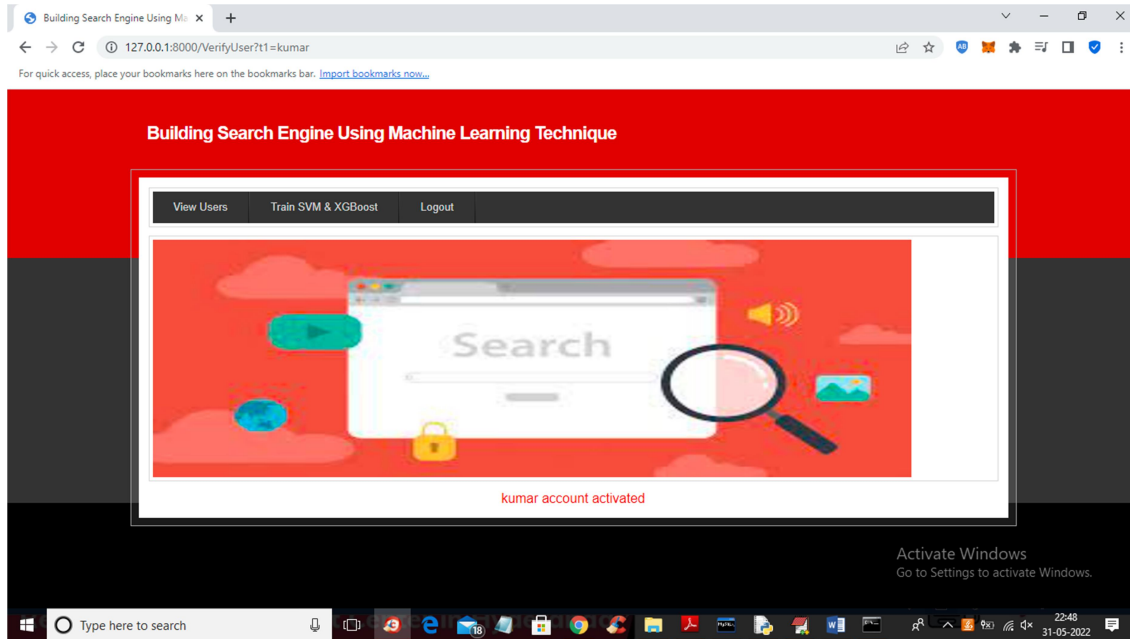
Username	Password	Contact No	Gender	Email Address	Address	Status
aaa	aaa	7778889998	Male	aaa@gmail.com	hyd	Accepted
kumar	kumar	6667778889	Male	kumar@gmail.com	hyd	Click Here

Activate Windows
Go to Settings to activate Windows.

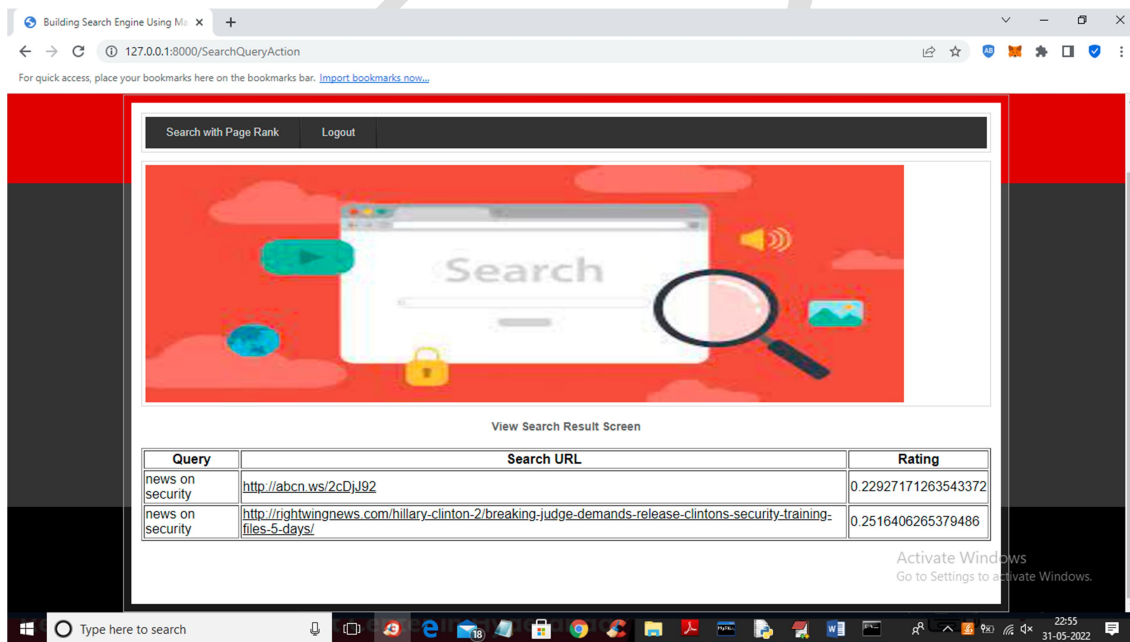
Type here to search

22:48
31-05-2022

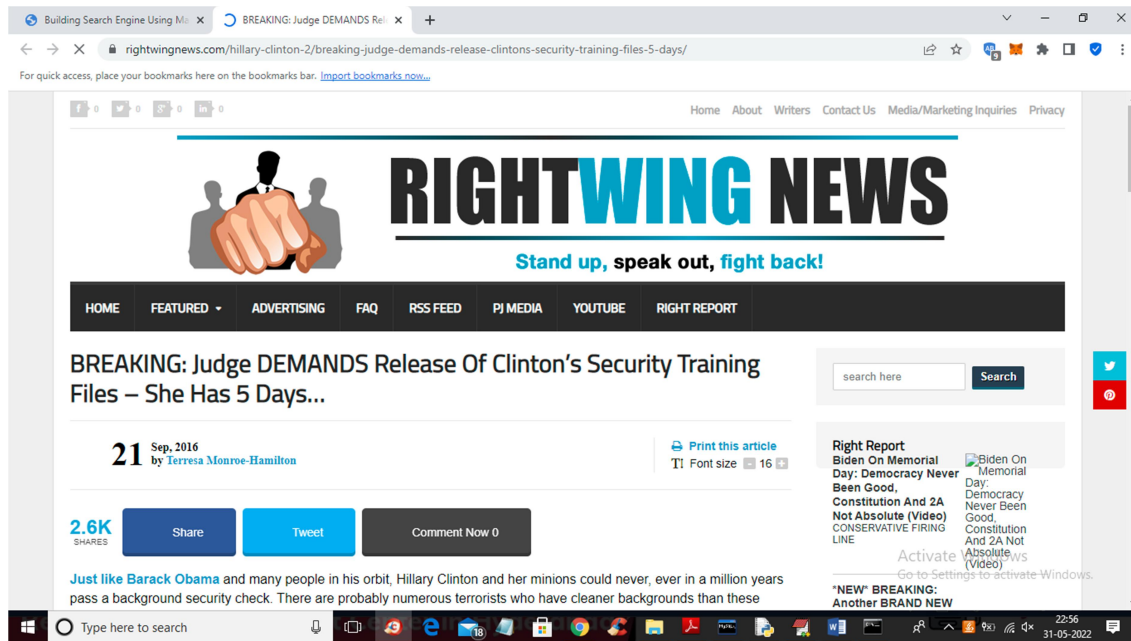
In above screen admin can click on 'Click Here' link to activate that user account



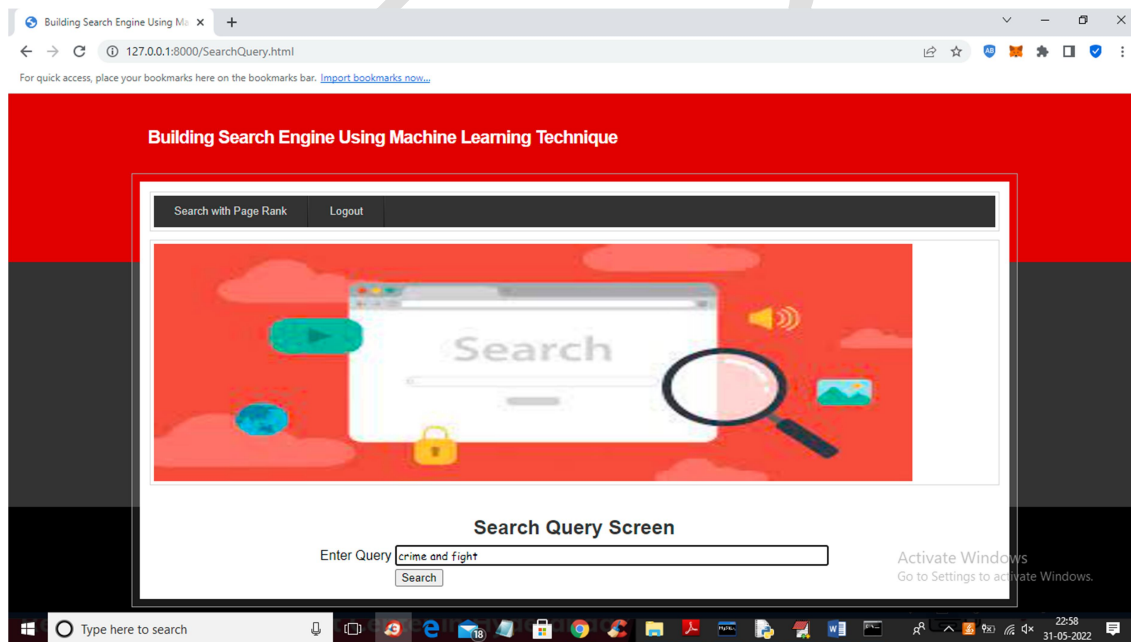
In above screen we can see admin activated kumar user account and now admin can click on ‘Train SVM & XGBOOST’ link to train machine learning SVM and XGBOOST algorithm and get below output



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



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



For above query we got below result

5.CONCLUSION

Search engines are quite effective for discovering more relevant URLs for specific keywords. As a result, the amount of time users spend searching for relevant web pages is reduced. Accuracy is quite crucial in this regard.

Based on the observations above, it is possible to conclude that XGBoost outperforms SVM and ANN in terms of accuracy. Thus, search engines developed with the XGBoost and PageRank algorithms will provide greater 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 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), Dec 2017.
- [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.

Author's Profiles

Dr. R. Konda Reddy working as a Professor in the Department of CSE, PBR Visvodaya Institute of Technology & Science, Kavali. He awarded Ph.D from Rayalaseema university, Kurnool in 2020. He has 22 years of teaching experience in various Engineering Colleges. He has published 27 research papers in national and international journals. His areas of interest include computer networks, MANET Routing with Intrusion Detection.



G. Geetha Madhuri, B. Tech with Specialization of Computer Science and Engineering in PBR VISVODAYA INSTITUTE OF TECHNOLOGY & SCIENCE (AUTONOMOUS), Kavali.



Y. Hima Chandana B. Tech with Specialization of Computer Science and Engineering in

PBR VISVODAYA INSTITUTE OF TECHNOLOGY & SCIENCE (AUTONOMOUS), Kavali.



N. Manoj, B. Tech with Specialization of Computer Science and Engineering in
PBR VISVODAYA INSTITUTE OF TECHNOLOGY & SCIENCE (AUTONOMOUS), Kavali.



Sk. Ameer Basha, B. Tech with Specialization of Computer Science and Engineering in
PBR VISVODAYA INSTITUTE OF TECHNOLOGY & SCIENCE (AUTONOMOUS), Kavali.