Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

Detection of Phishing Websites using an Efficient Machine Learning Framework

Naresh Kumar D
Dept. of IT
Panimalar Engineering College
Chennai, India

Nemala Sai Rama Hemanth
Dept. of IT
Panimalar Engineering College
Chennai, India

Premnath S
Dept. of IT
Panimalar Engineering College
Chennai, India

Nishanth Kumar V
Dept. of IT
Panimalar Engineering College
Chennai, India

Uma S
Dept. of IT
Panimalar Engineering College
Chennai, India

***Abstract*—Phishing attack is one of the commonly known attack where the information from the internet users are stolen by the intruder. The internet users are losses their sensitive information such as Protected passwords, personal information and their transactions to the intruders. The Phishing attack is normally carried by the attackers where the legitimate frequently used websites are manipulated and masked to gather the personal information of the users. The Intruders use the personal information and can manipulate the transactions and get definite from them. From the literature there are various anti-Phishing websites by the various authors. Some of the techniques are Blacklist or Whitelist and heuristic and visual similarity based methods. In spite of the users using these techniques most of the users are getting attacked by the intruders by means of Phishing to gather their sensitive information. A novel Machine Learning based classification algorithm has been proposed in this paper which uses heuristic features where feature selection can be extracted from the attributes such as Uniform Resource Locator, Source Code, Session, Type of security involve, Protocol used, type of website. The proposed model has been evaluated using five machine learning algorithms such as random forest, K Nearest Neighbor, Decision Tree, Support Vector Machine, Logistic regression. Out of these models, the random forest algorithm performs better with attack detection accuracy of 91.4%. Moreover the Random Forest Model uses orthogonal and oblique classifiers to select the best classifiers for accurate detection of Phishing attacks in the websites.**

***Keywords*—Phishing attack, Machine Learning, Classification Algorithms, Cyber Security, Heuristic Approach.**

I. INTRODUCTION

In this digital era, the people get interconnected with each other by means of internet with the help of the electronic devices like computers, laptops and PDA. Due to the revolution of the internet the most of the e-banking and e-commerce shopping had been preferred by most of their users due to his comfortness, availability and ease of use. Since, all these transactions or communications takes place in an open channel which is not secure in nature. The attacker tries to gain control over the insecure system which can cause various types of attacks during the transactions of the users. Phishing attack is one such type of attack where intruders tries to steal the user's sensitive and personal information by replicating the trustworthy websites to redirect the link to the intruder. In this Phishing attack the intruder tries to trap the legitimate user by generating the trustworthy webpage as a fraudulent webpage which is controlled by the attacker. Once when the legitimate user gives the personal information to the fraudulent website, their information is get recorded by the attacker from the background. By doing so all the sensitive information can be collected by the attacker by using phishing attack.

There are various types of Phishing attacks which has been used by the attackers in various domains for the different purpose. The mostly attacked domain for phishing attack is banking sector. In this domain the Phishing attack is normally occur when the user authenticates to the net banking using their username and password. At this point of time the attackers create the replication of both URL and webpage to make the user enter their credentials in the replicated fraudulent websites. By doing so the Credentials of the users getting recorded and they can gain access control to the user account without his concern. The next category of the phishing attack normally attacks in e-commerce websites. The intruders create the replica of the legitimate websites and make the users to carry out their transaction in the fake website. Once the Transaction are carried out the attackers record their credentials like username password and transaction parameters like ATM card number, pin number, and CVV number. Hence by caring this activities the attacker gain control over the system and can carry out the transaction on behalf of the legitimate user without his/her concern. These are such type of scenarios where phishing attack can cause harm to the legitimate users.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

In order to monitor the various phishing attack occur across the globe, a non-profitable Anti-Phishing Working Group is formed where the detailed investigation of various phishing attack are carried out and published in order to reveal malicious websites to the users. Normally the attackers create the fraudulent webpages and share to the users in forms of links through the social networking like Facebook, Instagram, WhatsApp and LinkedIn. As soon as link is clicked the users are directed to the fraudulent websites which can record their personal information. Current Phishing attacks are very powerful even the security services of various protocols like HTTPS, SSL can be breached. Hence the existing security mechanism are no longer secure. In order to overcome the limitations of existing systems in this paper can novel Phishing detection mechanism is proposed which is based on machine learning based classification to detect the phishing websites from the legitimate websites, more over the proposed method uses the URL based attributes as the input for the machine learning based classification algorithm by doing so the proposed method can successfully detect the normal websites from the fraudulent website and can control online phishing attack for the users in the internet.

II. LITERATURE SURVEY

Recently, Internet has become part of human lives. The current internet based Information and Communication Technology (ICT) prone to various threats and attacks which leads to significant loss. The basic goal of cyber security is to develop a security model to detect and prevent from the attacks. Various authors Selvi et al. (2019), Nancy et al. (2020), Rakesh et al. (2019), Santhosh Kumar et al. (2018), Thangaramya et al. (2020) shared their views on security in various fields.

Among them, Patrick Lawson et al. (2020) investigated the interaction between targeted user and persuation principle used in the domain of email phishing attack. They predicted vulnerabilities in phishing emails by using signal detection framework. Gonzalo De La Torre Parra et al. (2020) proposed framework for cloud based distributed environment for detecting phishing attack and botnet attack in Internet of things (IoT). They developed two security mechanism namely a Distributed Convolutional Neural Network (CNN) to detect phishing and Distributed Denial of Service (DDoS) attack and a cloud-based temporal Long-Short Term Memory for detecting botnet attacks. Their distributed CNN model were embedded with machine learning engine in the users IoT device.

Spear phishing attack is an attack where the attacker collects the user information on a specific victim profile or group of victim profile. Therefore, Luca Allodi et al. (2020) proposed new anti-phishing measure to protect legitimate user from spear phishing attack. Rui Chen et al. (2020) examines the effect of recent phishing and they focused process and outcome of Phishing detection and also they introduced deception theory to describe how the legitimate users experienced the difficulty in detection process and the outcomes have an impact on perceived susceptibility on phishing attack.

Justinas Rastenis et.al. (2020) broadly classified e-mail based phishing attack includes six stages of attack. Each stage has at least one measure to categorize the attacks. Each categorize have sub-section to explain the all variety of phishing attacks. They compared their proposed taxonomy with other similar taxonomies and identified their taxonomy performs well in terms of number of stages, measures and distinguished sections. Sahoo (2018) used a data mining technique to analyse phishing attacks on e-mail and built an architecture model separate regular e-mail from spam mail by using Naïve Bayes classification technique Sridharan and Sivakumar (2018), Sridharan and Chitra(2016), Sridharan and Chitra(2014). Niu et al, (2017) proposed a model to detect the phishing e-mails using the heuristic method based machine learning algorithm called Cuckoo Search-Support Vector Machine. This method extracts 23 features used to construct a hybrid classifier to optimize the feature selection of radial basis function.

M. Baykara and Z. Z. Gürel (2018), developed anti-phishing simulator, which provides information on the detection problem of a phishing attack and explained how to detect the phishing attack. This software examines mail content and identified phishing emails and spam emails. Şentürk et al. (2017) proposed an anti-phishing solution using machine learning and data mining technique to guard the user credentials against various attacks, namely spoofed emails and fraudulent websites. Mamoona Humayun et al. (2020) studied to identify and analyse the cyber security threats and vulnerabilities. They have identified how frequently the attacks occurred and also determine security vulnerabilities such as phishing, malware and Denial of Service. In spite of all these works many challenges needs to be addressed. Therefore, we proposed an efficient model to detect phishing attacks using various machine learning algorithms.

III. PROPOSED SYSTEM MODEL

The proposed system consists of two phases namely, Classification phase and phishing detection phase.

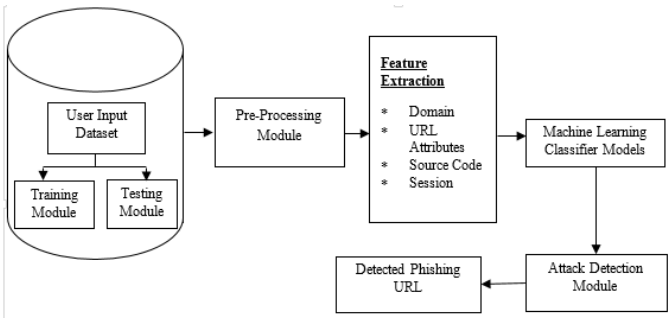

Fig. 1. Proposed Model to detect Phishing Attack

Fig. 1 gives the details of various steps carried out in classification of normal URL's with suspected phishing URL's.

A. Classification Phase

In the classification phase the input is URL's which comprises of both normal URL's and suspected Phishing website URL's. These inputs are given to three sub modules namely, Data Collection module, Feature selection module, classification module. In data collection module, the two types of URL's are considered, one is Phishing URL and another one is Legitimate URL's. From the data collection

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

module, the phishing URL's and Legitimate URL's are given feature extraction module. In feature extraction module it considers the attributes such as Address Bar, abnormal based feature, HTML and JavaScript and domain based feature. These attributes are given as an input to the classification module. The main goal of the classification module is to detect the phishing websites accurately from the normal URL's to the Phishing URL's. The main aim of the feature selection is to extract the valid and necessary features so that classifier is accurate in detecting the phishing URL's from the attributes given by the feature selection module. The proposed work comprises of five machine learning classifiers namely, K Nearest Neighbour (KNN), Decision Tree, Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM).

B. Phishing URL's Detection Module

The main aim of this module is to detect the legitimate URL's from the Phishing URL's based on attributes extracted in feature extraction module. Fig. 2 shows the phishing URL's detection module. In this module, the phishing URL's are given as a dataset.

The dataset is further divided into training dataset and testing dataset. The training dataset comprises of 70% and testing data set is comprised for 30%. The proposed module comprises of five machine learning classifiers namely, K Nearest-Neighbor (KNN), Logistic Regression (LR), Random Forest (RF), Decision Tree and Support Vector Machine (SVM).

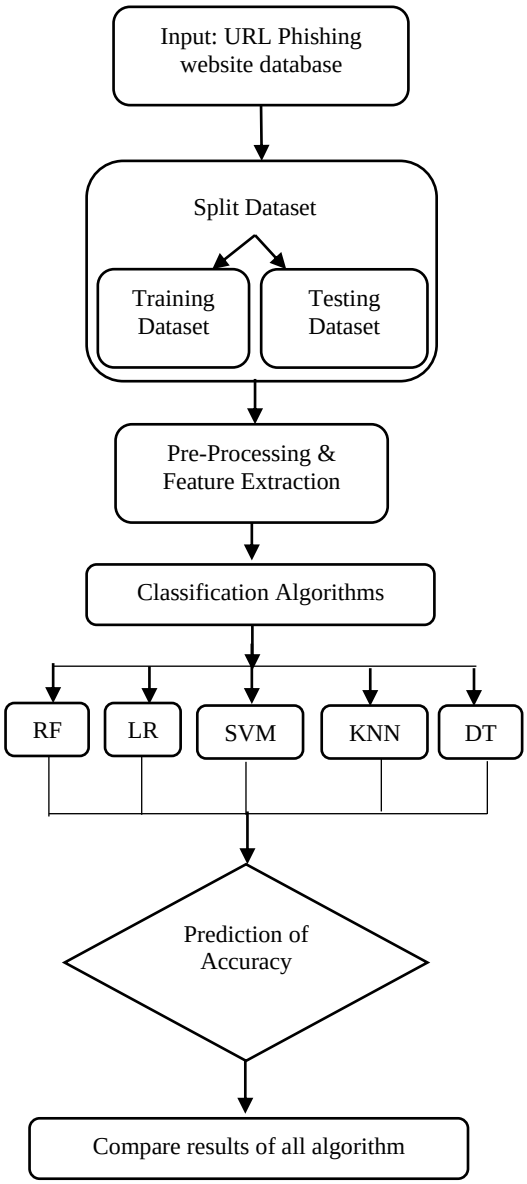

Fig. 2. Phishing URL detection module

1) K Nearest-Neighbour: The first machine learning classifier is K Nearest Neighbour. The K-nearest-Algorithm calculates the distance based on dataset and query scenario. The distance between the two points (x1,..,xn) and (y1,…,yn) are calculate based on the Euclidian distance. Based on the distance calculation, if the distance value is very less, (K-nearest-neighbour) then it considered as the phishing URL more over it ignores the other attributes in the data when the computed distance is more.

2) Decision Tree: The next category of machine learning classifier is decision tree algorithm. In decision tree the attributes with high information gain considered as different set of attributes where the certain decision can be obtain from the attributes of high information gain. In decision tree algorithm, the various phishing attributes with high information gain are compared with each other, the phishing attributes with high impact are considered as Phishing URL's and rest of the attributes are considered as legitimate URL's.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

3) Logistic Regression: The logistic regression is a kind of predictive analysis where based on the attributes the phishing URL's can be detected. In logistic regression the input is given as training data and testing data. Based on the given input logistic regression is computed by using the regression function called sigmoid function with the computed sigmoid function the relationship between training data and testing data is calculated. Based on the relation the objects are categorized. If the patterns in the attributes of the training data and testing data are same, then the URL's are considered as phishing URL's else other URL's are considered as Legitimate URL's.

4) Random Forest: The next category of machine learning is random forest algorithm. The main aim of the random forest is to detect the phishing URL's from the legitimate URL's. Random forest is widely used ensemble learning methods and works by combination of all their output and predicts the best output among the test data. They computes the Gini index method at each separation and uses the best split to provide the output. Random forest aggregates family classifier h(x|θ1),h(x|θ2),..h(x|θk) ,here h(x|θ), is a classification tree and k is the number of trees chosen from random vector model. Each θk is a randomly chosen parameter vector. D(x,y) indicates the training dataset, each classification tree in the ensemble is built using a different subset Dθk(x,y) ⊂ D(x,y) of the training dataset.

Thus, h(x|θk) is the kth classification tree which uses subset of features xθk ⊂ x to build a classification model. They are like normal decision tree.

The output of y shown in equation (1)

$$y = argmax_{p\in\{h(x_1)\ldots h(x_k)\}}\left\{\sum_{j=1}^{k}\left(I\left(h\left(x\middle|\theta_j\right)=p\right)\right)\right\} \quad (1)$$

IV. RESULTS AND DISCUSSIONS

The proposed model is evaluated by using python. We have considered 4 performance metrics namely, Root-Mean-Square Error (RMSE), R squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE).

Fig. 3 gives the Mean Square error value for the four Different Machine Learning classification Algorithm.

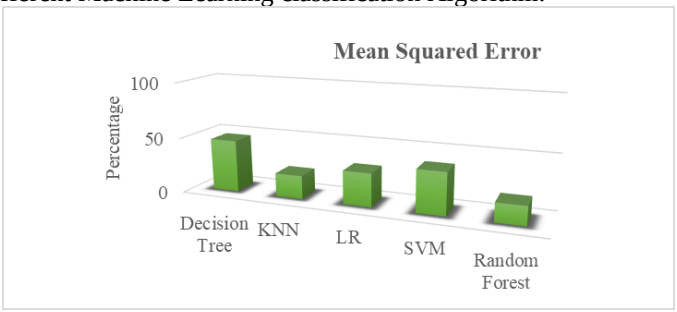

Fig. 3. Comparison analysis of classification algorithms for MSE

From the graph it is cleared that the random forest algorithm has better MSE value, when it is compared with other machine learning classifier algorithms. Since the random forest algorithm has least MSE value, it significantly increases the accuracy of Phishing attack detection.

Fig. 4 gives the R Squared value for the four Different Machine Learning classification Algorithm

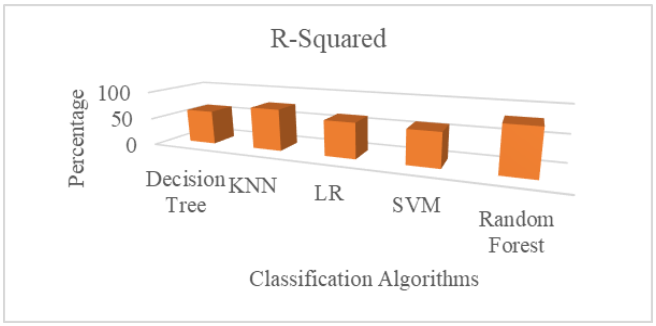

Fig. 4. Comparison Analysis of classification algorithms for R-Squared

From the graph it is cleared that the random forest algorithm has higher R-squared value, when it is compared with other machine learning classifier algorithms. Since the random forest algorithm has higher R-squared value, it significantly increases the accuracy of Phishing attack detection.

Fig. 5 gives the Mean Absolute Error (MAE) value for the four Different Machine Learning classification Algorithm.

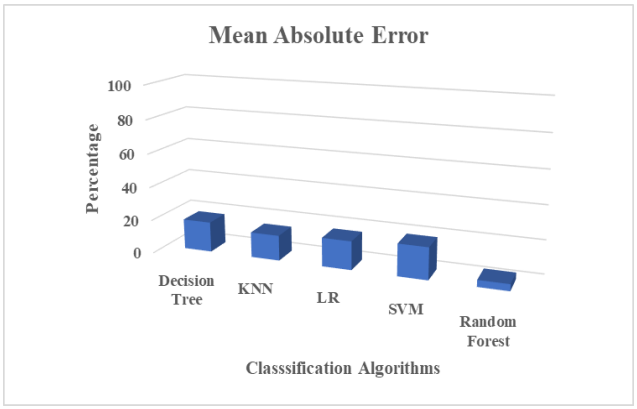

Fig. 5. Comparison analysis of classification algorithms for MAE

From the graph it is cleared that the random forest algorithm has least Mean Absolute Error value, when it is compared with other machine learning classifier algorithms. Since the random forest algorithm has least Mean Absolute Error value, it significantly increases the accuracy of Phishing attack detection.

Fig. 6. gives the Root Mean Squared Error (RMSE) value for the four Different Machine Learning classification Algorithm.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

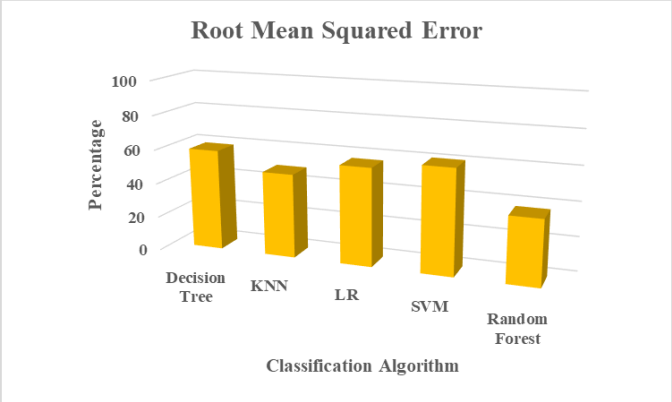

Fig. 6. Comparison Analysis of classification algorithms for RMSE

From the graph it is cleared that the random forest algorithm has least Root Mean Squared Error value, when it is compared with other machine learning classifier algorithms. Since the random forest algorithm has less Root Mean Squared Error value, it significantly increases the accuracy of Phishing attack detection

V. CONCLUSION AND FUTURE WORK

Phishing attack is one of the common type of cyber-crime where the attackers can steal the user's personal information by forgery the legitimate website with the masked one. The Proposed system uses five different machine Learning classifiers namely, Decision Tree, Random Forest, K-Nearest-Neighbor, Logistic Regression and Support Vector Machine. These algorithms are implemented by the Performance metrics like Root Mean Square Error (RMSE), R-Squared, Mean Absolute Error (MAE) and Mean Squared Error (MSE). From the experimental result is cleared that the random forest algorithm has higher R-Squared Value and better Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE). Moreover, the Random Forest classifier has better phishing detection accuracy of 91.4% compared with other machine learning classifier. The future work of the proposed system is to evaluate these machine learning classifiers with larger dataset.

REFERENCES

[1] Patrick Lawson, Carl J. Pearson, Aaron Crowson, Christopher B. Mayhorn, "Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy", Applied Ergonomics, Elsevier, vol. 86, pp. 1-10, 2020.

[2] Gonzalo De La Torre Parra , Paul Rad, Kim-Kwang Raymond Choo, Nicole Beebe, "Detecting Internet of Things attacks using distributed deep learning", Journal of Network and Computer Applications, Elsevier, vol. 163, pp. 1-13, 2020.

[3] Luca Allodi, Tzouliano Chotza, Ekaterina Panina, and Nicola Zannone, "The Need for New Antiphishing Measures Against Spear-Phishing Attacks", IEEE Security & Privacy, pp. 23-34,2020.

[4] Rui Chen, Joana Gaia, H. Raghav Rao, "An examination of the effect of recent phishing encounters on phishing susceptibility", Elsevier, pp.1-14, 2020.

[5] Justinas Rastenis, Simona Ramanauskaite, Justinas Janulevicius , Antanas Cenys , Asta Slotkiene and Kestutis Pakrijauskas, "E-mail-Based Phishing Attack Taxonomy", Applied Sciences, MDPI, vol. 10, pp.1-15, 2020.

[6] P. K. Sahoo, "Data mining a way to solve Phishing Attacks," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), IEEE, pp. 1-5, 2018.

[7] W. Niu, X. Zhang, G. Yang, Z. Ma and Z. Zhuo, "Phishing Emails Detection Using CS-SVM," International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), Guangzhou, IEEE , pp. 1054-1059, 2017.

[8] M. Baykara and Z. Z. Gürel, "Detection of phishing attacks," 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, pp. 1-5, 2018.

[9] Ş. Şentürk, E. Yerli and İ. Soğukpınar, "Email phishing detection and prevention by using data mining techniques," International Conference on Computer Science and Engineering (UBMK), Antalya, pp. 707-712, 2017.

[10] Mamoona Humayun, Mahmood Niazi, NZ Jhanjhi, Mohammad Alshayeb and Sajjad Mahmood, "Cyber Security Threats and Vulnerabilities: A Systematic Mapping Study", Arabian Journal for Science and Engineering, Springer, vol. 45, pp. 3171–3189, 2020.

[11] M Selvi, K Thangaramya, Ganapathy Sannasi, K Kulothungan, H Khannah Nehemiah, A. Kannan, "An Energy Aware Trust Based Secure Routing Algorithm for Effective Communication in Wireless Sensor Networks", Wireless Personal Communications, Springer, pp1-16, 2019.

[12] Periasamy Nancy, Sannasy Muthurajkumar, Sannasi Ganapathy, S. V. N. Santhosh Kumar, M. Selvi, Kannan Arputharaj, "Intrusion detection using dynamic feature selection and fuzzy temporal decision tree classification for wireless sensor networks", IET Communications, vol.14, pp.888-895, 2020.

[13] Rakesh Rajendran, S. V. N. Santhosh Kumar, Yogesh Palanichamy, Kannan Arputharaj, "Detection of DoS attacks in cloud networks using intelligent rule based classification system", Cluster Computing, vol.22 pp.423-434, 2019.

[14] S. V. N. Santhosh Kumar, Yogesh Palanichamy, "Energy efficient and secured distributed data dissemination using hop by hop authentication in WSN", Wireless Networks, vol.24, pp.1343-1360, 2018.

[15] K Thangaramya, K Kulothungan, S Indira Gandhi, M Selvi, SVN Santhosh Kumar, Kannan Arputharaj," Intelligent fuzzy rule-based approach with outlier detection for secured routing in WSN", Soft Computing, Springer Berlin Heidelberg, pp.1-15, 2020.

[16] K. Sridharan and P. Sivakumar, "Hybrid Approach Analysis for Text Categorization Using Intuitive Classifiers", Journal of Computational and Theoretical Nanoscience, vol. 15, pp. 811-822, 2018.

[17] K. Sridharan and M. Chitra, "Experimental Investigation for Text Categorization Based on Hybrid Approach Using Feature Selection and Classification Techniques", Asian Journal of Information Technology, vol. 14, no.15, pp.2355 - 2366, 2016.

[18] K Sridharan and M. Chitra, "Trust based automatic query formulation search on expert and knowledge users systems", Journal of Computer Science, vol. 10, pp. 1174-1184, 2014.