International Journal of Computer Applications Technology and Research
Volume 5– Issue 4, 176 - 182, 2016, ISSN:- 2319–8656

Rice Seed Germination Analysis

Benjamaporn Lurstwut
Faculty of Information
and Communication Technology,
Mahidol University, Thailand

Chomtip Pornpanomchai
Faculty of Information
and Communication Technology,
Mahidol University, Thailand

Abstract: This research aimed to develop the computer software called "Rice Seed Germination Analysis (RiSGA)" which could predict rice seed image for rice germination by using an image processing technique. The RiSGA consisted of five main process modules: 1) image acquisition, 2) image pre-processing, 3) feature extraction, 4) quality control analysis and 5) quality results. Six variations of Thai rice seed species (CP111, RD41, Chiang Phattalung, Sang Yod Phattalung, Phitsanulok 2 and Chai Nat 1) were used for the experiment. The RiSGA extracted three main features: 1) color, 2) morphological and 3) texture feature. The RiSGA applied four well-known techniques: 1) Euclidean Distance (ED), 2) Rule Based System (RBS), 3) Fuzzy Logic (FL) and 4) Artificial Neural Network (ANN). The RiSGA precision of ED, RBS, FL, and ANN was 87.50%, 100%, 100%, and 100%, respectively. The average access time was 4.35 seconds per image, 5.29 seconds per image, 7.04 seconds per image, and 159.65 seconds per image, respectively.

Keywords: Rice seed, seed germination, rice seed features, image processing, computer vision

1. INTRODUCTION

Nowadays, the agricultural industry is more widespread in the world. Oryza Sativa (Rice) is a vital worldwide agricultural produce which is very popular [47]. Thailand is one of agricultural countries which produces a large number of food products e.g. cereals, flowers, vegetables, fruits, rubber, and especially rice. In Thailand, there are more than 114 well-known Thai rice species [33, 39, 46]. Paddy rice based on the good quality of products is mostly offered [40]. Therefore, it is essential to grade the quality of these commodities in order to command the better price in the market competition. The quality of the rice is mostly based on the quality of the seeds. However, it is very difficult to identify the quality of rice seeds by using only human vision. Thus, this research aims to apply the standard seed germination test from the International Seed Testing Association (ISTA 1996) with the top of paper method for rice germination [15, 16, 22, 23, 31]. Moreover, images are collected to predict rice seed images by using image processing techniques [18] which can identify the quality of products. The objective of this research is to develop the computer software which can predict rice seed microscopic image for rice germination by using image processing techniques. This research focuses on six variations of Thai rice seed species including CP111, RD41, Chiang Phatthalung, Sang Yod Phattalung, Phitsanulok 2 and Chai Nat 1.

2. RELATED WORKS

Many researchers [1, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 17, 19, 20, 21, 24, 26, 27, 29, 30, 32, 34, 35, 36, 37, 38, 41, 42, 43, 44, 45, 48, 49, 50, 51] have applied image processing for rice germination in their studies. This is because using image processing is rapid, economic, consistent, and objective.

However, there is no research conducted by using machine vision and image processing in the microscopic level which is more efficient for analyzing the information. Therefore, this research aims to implement the automatic system in the microscopic level. The Rice Seed Germination Analysis (RiSGA) system is developed and used as the inspection tool for measuring the rice seed quality in the rice germination which is essential to grade the quality of products in order to command the better price in the market competition.

3. MATERIALS AND METHODS

The experiment was conducted by using the following computer hardware specifications: 1) CPU Intel(R) Core(TM) i5-2400 CPU @ 3.10GHz 2) Memory DDR3 4 GB and 3) Hard disk 500 GB. For the computer software, Microsoft Windows 7 (Microsoft Corp.; Redmond, WA, USA) was used as the operating system. For the development tool, MATLAB R2013a (The Math Works Inc.; Natick, MA, USA) [28] was used.

Analysis and design were described by using the system conceptual diagram and system structure chart. The details of each element are described below.

3.1 System conceptual diagram

The RiSGA system is a system to imitate the abilities of the classification of the quality control of the seeds. Fig. 1 shows the overview of the system conceptual diagram. The operation of the RiSGA system is divided into two phases: 1) the preparation of the system data set for training data set and testing data set, and 2) the preparation of the system data set for testing the system on another data set of unseen or unknown images. The two phases has to be operated in order to observe its performance and accuracy. However, the second phase has to be used to validate and evaluate the RiSGA system efficiently.

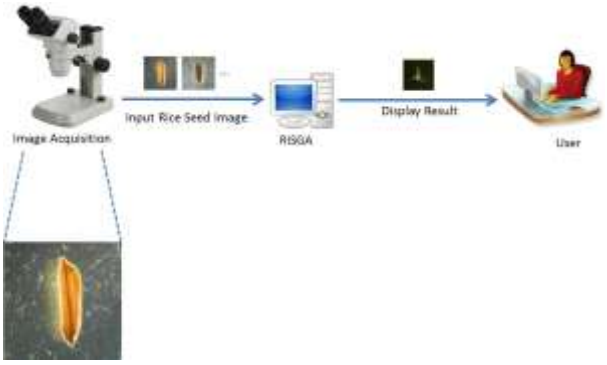


Figure. 1 The overview of the system architecture of the RiSGA system.

3.2 System Structure Chart

The RiSGA structure chart elaborates on how each model works is shown in Fig. 2. The RiSGA consists of five main process modules: 1) image acquisition, 2) image pre-processing, 3) feature extraction, 4) quality control analysis and 5) quality results. Each process module has the following details.

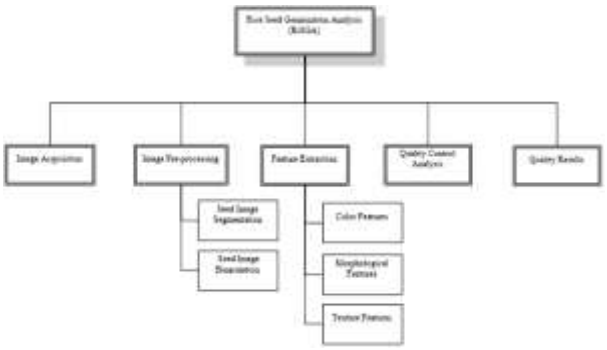


Figure. 2 Rice Seed Germination Analysis System Structure Chart.

3.2.1 Image acquisition

In this module, it consists of three main parts which are capturing image, cropping microscopic image, and storing image. In capturing the image, the rice seed is placed on the plate and then the rice seed is taken in a bird-eyes-view angle from a digital microscope camera on a light mode or a simple camera. The rice seed image will be used as the input to the RiSGA system. In cropping the image, the size of each image captured from the microscope camera and the simple camera is very large and there is a lot of noise. Therefore, the raw microscopic image must be cropped in the same scale of 1900 x 1900 pixels at the same position. In storing image, the rice seed images from the microscope camera and the digital camera are collected and stored in the system database.

3.2.2 Image Pre-processing

In the second module, the image pre-processing module prepares an image before processing in the feature extraction process. This module consists of two sub-modules which are seed image segmentation and seed image binarization.

3.2.2.1 Seed Image Segmentation

The sample of rice seed microscopic image segmentation is shown in Figure. 3. Figure. 3 (a) shows a rice seed microscopic image. Figure. 3 (b) shows a bounding box of rice seed microscopic image. Figure. 3 (c) shows an enhanced bounding box of rice seed microscopic image. Figure. 3 (d) shows a segmented rice seed microscopic image.

First, the RiSGA system finds the bounding box of the rice seed microscopic input image by changing the RGB rice input microscopic image to the binarization image (Figure. 3 (a) - Figure. 3 (b)). Next, the RiSGA system will perform morphological closing to close any opening area. The RiSGA system fills holes and removes noise in order to get the enhanced binarization image. Then, the RiSGA system labels the eight connected components of the enhanced binarization image to build the rectangle which can cover and fit the size of the rice seed object in order to be the label bounding box of the rice seed microscopic image (Figure. 3 (c)). Finally, the RiSGA system uses the label bounding box of the rice seed microscopic image to segment only the rice seed object in the input image (Figure. 3 (d)).

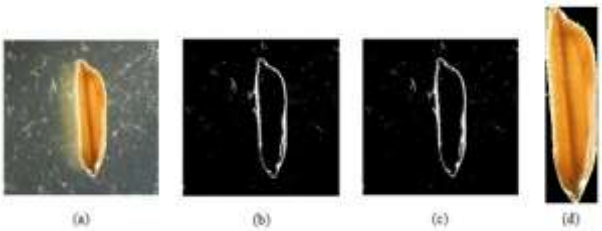


Figure. 3 The sample of rice seed microscopic image segmentation.

3.2.2.2 Seed Image Binarization

The sample of rice seed microscopic image binarization is shown in Figure. 4. Figure. 4 (a) shows a segmented rice seed microscopic image. Figure. 4 (b) shows a rice seed microscopic gray-scale image. Figure. 4 (c) shows a rice seed microscopic binarization image. Figure. 4 (d) shows an enhanced rice seed microscopic binarization image.

First, the RiSGA system changes the RGB color image to the gray-scale image (Figure. 4 (a) - Figure. 4 (b)). Next, the RiSGA system transforms the gray-scale image to the binary image (Figure. 4 (c)). Then, the RiSGA system fills holes and removes noise in order to get the enhanced binarization image (Figure. 4 (d)).

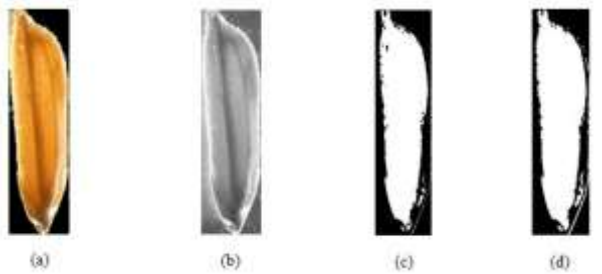


Figure. 4 The sample of rice seed microscopic image binarization.

3.2.3 Features extraction

In the third module, the feature extraction module is the most important module. This is because this module contains rice images which are cropped from Module 1. Moreover, this module is used to describe the accuracy of the large set of the data. In the classification process, there are three main sub-modules or seven combination optimal features as follows:

- Color feature consists of (1) average red color, (2) average green color, (3) average blue color

- Morphological feature consists of (4) aspect ratio, (5) edge,

- Texture feature consists of (6) entropy, (7) energy

To define the optimal features, the feature selection process is used to observe the most relevant features of the seed characteristics. In this research, the Wrapper Feature Selection Approach [25] is selected to search the optimal feature subset in order to achieve the best performance. After the Wrapper Feature Selection Approach had been applied, the three main sub-modules or the seven combination optimal features occurred which were (1) average red color, (2) average green color, (3) average blue color, (4) aspect ratio, (5) edge, (6) entropy, and (7) energy.

3.2.3.1 Color Feature

There are many color spaces in color features. In this research, the RGB color space is considered as the main color feature. The RGB feature consists of three features: (1) average read color, (2) average green color, and (3) average blue color. The three features are considered as the optimal selected color features.

3.2.3.2 Morphological Feature

The morphological feature measures the features based on the seed morphology, especially for the shape and the size of the seed. The morphological feature consists of two optimal features which are (1) aspect ratio and (2) edge. The aspect ratio is calculated by the major axis length divided by minor axis length. The edge is calculated by applying the Sobel edge detection with threshold values 0.03 to measure the remaining pixels in the seed area.

3.2.3.3 Texture Feature

The RiSGA system applies gray level co-occurrence matrices (GLCM) [2] for measuring the seed surface texture. The GLCM is extracted from each of the gray-tone spatial-dependence matrices. There are many features in the GLCM. However, the two optimal features: (1) energy and (2) entropy are applied in this research. Each texture feature is calculated based on equation (1) - (2):

Where

$P_{i,j}$ = entry in a normalized gray-tone spatial-dependence matrix,

N = number of distinct gray levels in the quantized image.

(1) Energy Texture Feature

The energy texture feature known as uniformity is the sum of squared elements in the GLCM. The energy texture is calculated in equation (1):

$$\mu_i = \sum_{i,j=0}^{N-1} iP_{i,j}, \mu_j = \sum_{i,j=0}^{N-1} jP_{i,j} \quad (1)$$

(2) Entropy Texture Feature

The entropy texture feature is a statistical measure of randomness which is used to characterize the texture of the input image. The entropy texture is calculated in equation (2):

$$\sum_{i,j=0}^{N-1} P_{i,j}\left(-\ln P_{i,j}\right) \quad (2)$$

3.2.4 Quality Control Process

In the fourth module, the RiSGA system uses the quality control analysis to apply four techniques to predict rice seed image for rice germination: (1) Euclidean Distance (ED), (2) Rule Based System (RBS), (3) Fuzzy Logic (FL) and (4) Artificial Neural Network (ANN). Next, the RiSGA system compares the features of the testing rice seed image data set with the training rice seed image data set in the system database.

3.2.4.1 Euclidean Distance (ED)

The Euclidean Distance (ED) is used to measure the similarity of the distance between every feature of a sample data set and every feature of each training data set in the RiSGA system. The RiSGA system applies the Euclidean Distance based on the minimum distance. The minimum Euclidean Distance determines that the sample data set and the training data set are very similar. The Euclidean Distance is calculated in equation (3):

$$ED = \sqrt{\sum_{i-1}^{n}\left(x_i - y_i\right)^2} \quad (3)$$

Where ED = the Euclidean distance value between two objects which are x and y, n = number of features, x_i = the value of feature i in the system database and y_i = the value of feature i in a sample image.

3.2.4.2 Rule Based System (RBS)

The rule based system is used to store and manipulate knowledge. The knowledge is stored as the rule based representation in IF-THEN structure. The fact is represented in the IF part (antecedent). The action is represented in the THEN part (consequent). The RiSGA system applies the rule based system from the area under the normal curve in the normal distribution of data set.

3.2.4.3 Fuzzy Logic (FL)

The fuzzy logic is determined as a set of mathematical principles for knowledge representation based on degree of membership function. To define the appropriate degree of

membership function, the RiSGA system applies the mean and the quantity of three times the standard deviation to define the degree of membership function. The feature values in the data set are normalized to the scale of 0 and 1 which is easier to determine the degree of the membership function. The RiSGA system applies the trapezoidal function for representing the degree of the membership function. The input range were between 0 and 1 while the output range were between -0.3 and 1.5.

3.2.4.4 Artificial Neural Network (ANN)

ANN is a mathematical model or computational model that is inspired by the structure and functional aspects of biological neural networks. ANN is composed of many artificial neurons that are linked together according to specific network architecture. The patterns feed the input into the network and then the network will return the output. The objective of the neural network is to transform the input into the meaningful output. In this research, the artificial neural network (ANN) classifies the rice images by using the neural network structure of 7-6-2. The seven input nodes are equal to seven features of each seed image and the two output nodes are equal to two kinds of germinated seed and non-germinated seed in the training data set. The hidden nodes are 2/3 of average between input nodes and output nodes which is the rule of thumb [5]. Figure. 5 shows the design of the neural network.

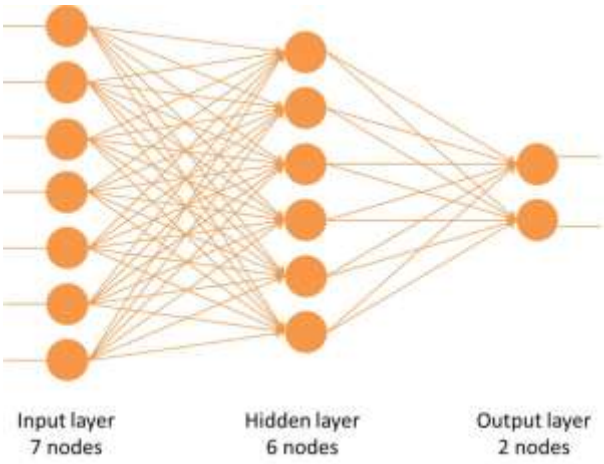


Figure. 5 The design of the neural network.

The neural network consists of input layer, hidden layer, and output layer. The input layer consists of 7 nodes. The hidden layer consists of 6 nodes. The output layer consists of 2 nodes.

3.2.5 Quality Result

The quality result module shows rice seed germinating prediction results displaying as the graphic user interface. This module is categorized into two parts which are Display Quality Result Image and Display Details.

3.2.5.1 Display Quality Result Image

Once the user load the unknown input image to the system, the system will extract the features of the unknown input image. Next, the system compared the features of unknown image with the system database. Then, the RiSGA system will analyze the result by using one of five techniques. Finally, the quality result image will display in the graphic user interface as shown in Figure. 6.

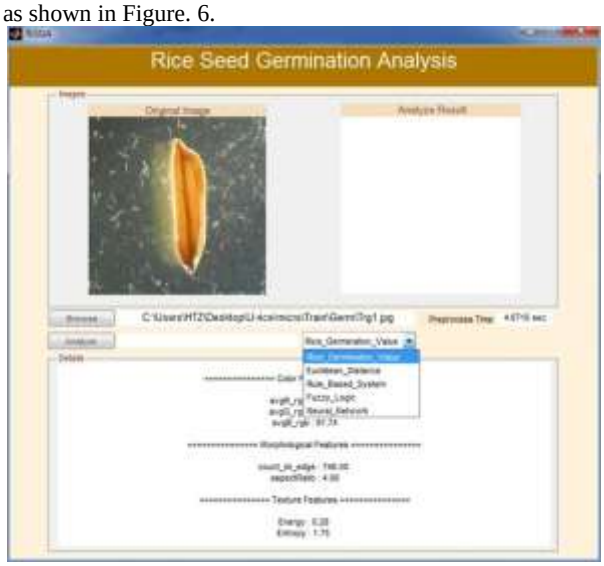


Figure. 6 The GUI of selecting the classification technique.

3.2.5.2 Display Details

The analyzing image result was displayed in the germination analysis result image box. The details of the analyzing image were displayed in the analyze result text box. Figure. 7 shows the GUI of displaying the quality control analysis result.

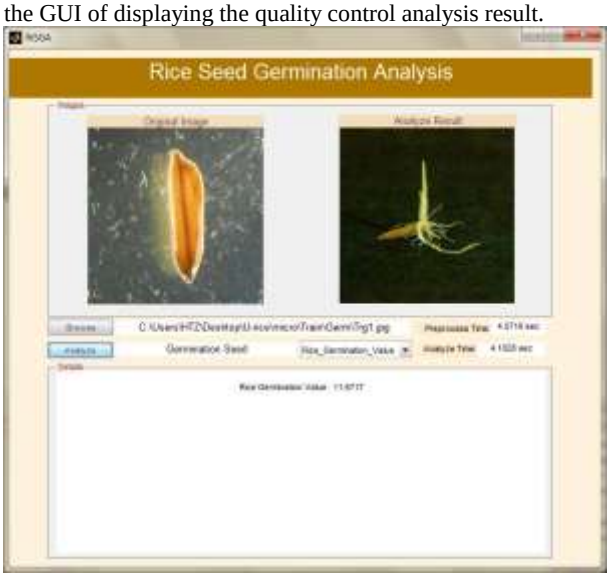


Figure. 7 The GUI of displaying the quality control analysis result.

4. RESULTS AND DISCUSSIONS

The RiSGA process employs six Thai rice seed species which are 1) CP111, 2) RD41, 3) Chiang Phatthalung, 4) Sang Yod Phattalung, 5) Phisanulok 2 and 6) Chai Nat 1 from Bureau of Rice Seed, Ministry of Agriculture and Cooperatives, Rice Department of Thailand. The sample of rice seed germination testing based on the top of paper method. For the evaluation of the five learning algorithms, each kind of rice species contains 600 samples in the microscope training data set and 120 samples in the microscope unknown data set.

4.1 Experiment Result based on training data set

The experiment result based on training data set is shown in Table 1.

Table 1. Experiment Result based on training data set

TQ	Amt	G		NG		A (%)	T (s/img)
		GP	N-GP	GP	N-GP		
ED	600	360	0	0	240	100.00	4.36
RBS	600	348	12	11	229	96.17	5.39
FL	600	360	0	0	240	100.00	6.08
ANN	600	360	0	0	240	100.00	162.05

Where TQ = Technique, Amt = Amount, G = Germination, NG = Non-Germination, GP = Germinated Prediction, N-GP = Non-Germinated Prediction, A = Accuracy, T = Time, and s/img = seconds per image

4.2 Experiment Result on unknown data set

The experiment result based on unknown data set is shown in Table 2.

Table 2. Experiment Result based on unknown data set

TQ	Amt	G		NG		A (%)	T (s/img)
		GP	N-GP	GP	N-GP		
ED	120	68	0	15	37	87.50	4.35
RBS	120	68	0	0	52	100.00	5.29
FL	120	68	0	0	52	100.00	7.04
ANN	120	68	0	0	52	100.00	159.65

Where TQ = Technique, Amt = Amount, G = Germination, NG = Non-Germination, GP = Germinated Prediction, N-GP = Non-Germinated Prediction, A = Accuracy, T = Time, and s/img = seconds per image

5. CONCLUSIONS

In this research, the Rice Seed Germination Analysis (RiSGA) system developed from MATLAB R2013a running on Windows 7 Ultimate and the Wrapper Feature Selection Approach were selected for performing feature selection. Later, the system was developed to predict rice germination in the seed quality control process. Six Thai rice seed species which were CP111, RD41, Chiang Phatthalung, Sang Yod Phattalung, Phisanulok 2, and Chai Nat collected from Bureau of Rice Seed, the Ministry of Agriculture and Cooperatives, and the Rice Department of Thailand were selected for the experiment. The RiSGA data set which consisted of 720 microscopic rice seed images were used for training data set and unknown data set. The germination experiment was conducted under the standard of the International Seed Test Association (ISTA) 1996. The environment of germination test was in the temperature of 20-30°C and the relative humidity was 72% only. The time of the experiment was from September 2013 to October 2014. Each germination tray was tested within 2 weeks. The system was applied four techniques for quality control process. The accuracy based on training data set of ED, RBS, FL, and ANN was 100%, 96.17%, 100% and 100% respectively. The average access time based on ED, RBS, FL, and ANN was 4.36, 5.39, 6.08, and 162.05 seconds per image, respectively. The accuracy based on unknown data set of ED, RBS, FL, and ANN was 87.50%, 100%, 100% and 100% respectively. The average access time based on ED, RBS, FL, and ANN was 4.35, 5.29, 7.04, and 159.65 seconds per image, respectively.

6. RECOMMENDATIONS

According to the experiment results, it was proved that the RiSGA system was efficient, effective, accurate, upgradable, and objective. However, the RiSGA system was used for the experiment for the only six Thai rice seed species collected in Thailand. Therefore, the RiSGA system should be used for the experiment for more data set in order to cover all rice seed species not only for rice seeds in Thailand but also for rice seeds in all parts of the world. To increase the efficiency of the system, the RiSGA system can be applied other feature selection to observe the key important features which are suitable for the rice seed characteristics in the rice germination, can be enhanced the Data Set to support not only Thai rice seed species but also all kinds of the rice seed species by increasing the variation of rice seed species in system database, can be applied other substratum in ISTA to enhance the ability of processing rice seed quality control process, can be improved image acquisition for supporting mobile phone devices which can be further developed for the online system application.

7. ACKNOWLEDGMENTS

This research is supported by Bureau of Rice Seed, Ministry of Agriculture and Cooperatives; Rice Department of Thailand [6] and Faculty of Science, Mahidol University for supplying rice seed species and facilities to the research.

8. REFERENCES

[1] Ajay, G., Suneel, M., Kumar, K. K. & Prasad, P. S., 2013, Quality Evaluation of Rice Grains Using Morphological Methods. In International Journal of Soft Computing and Engineering (IJSCE)., ISNN: 2231-2307., Vol. 2. Issue 6.

[2] Beliakov, G., James, S. & Troiano, L., 2008, Texture recognition by using GLCM and various aggregation functions. In Proceeding of the International Conference on Fuzzy System., pp.1472-1476, Hong Kong, China.

[3] Belsare, P. P. & Dewasthale, M. M., 2013, Application Of Image Processing For Seed Quality Assessment: A Survey. In International Journal of Engineering Research & Technology (IJERT)., ISSN: 2278-0181., Vol. 2. Issue 2.

[4] Belsare, P. P. & Shah, S. K., 2013, Evaluating of Seedling Growth Rate using Image Processing. In IEEE International Conference on Computational Intelligent and Computing Research (ICCIC)., pp. 1-4.

[5] Boger, Z., 1997, Knowledge Extraction from Artificial Neural Networks Models In Systems, Man, and Cybernetics, 1997. Computational Cybernetics and

Simulation., 1997 IEEE International Conference., ISSN: 1062-922X., Vol. 4., pp. 3030-3035. Orlando, FL.

[6] Bureau of Rice Seed [homepage on the Internet]. Thailand: Associate Online Resources. Retrieved November 17, 2013 available from: http://brs.ricethailand.go.th/

[7] Chaugule, A., 2012, Application of image processing in seed technology: A survey. In International Journal of Emerging Technology and Advanced Engineering., ISSN: 2250-2459., Vol. 2. Issue 4.

[8] Dell'Aquila, A., 2004, Application of a Computer-Aided Image Analysis System to Evaluate Seed Germination under Different Environmental Conditions. In Italian Journal of Agronomy., Vol. 8. pp. 51-62.

[9] Dell'Aquila, A., 2006, Computerised seed imaging: a new tool to evaluate germination quality. In Communications in Biometry and Crop Science. International Journal of the Faculty of Agriculture and Biology. Warsaw Agricultural University., Vol. 1. No. 1. pp. 20-31., Poland.

[10] Dell'Aquila, A., 2009, Digital Imaging Information Technology Applied to Seed Germination Testing: A Review. In Agronomy for Sustainable Development. International Journal., ISSN: 1773-0155., Vol. 29. Issue 1. pp. 213-221., Netherlands.

[11] Dell'Aquila, A., 2009, New Perspectives for Seed Germination Testing Through Digital Imaging Technology. In The Open Agriculture Journal. pp. 37-42.

[12] Ducournau, S., Feutry, A., Plainchault, P., Revollon, P., Vigouroux, B. & Wagner, M. H.. 2004, An image acquisition system for automated monitoring of the germination rate of sunflower seeds. In Computers and Electronics in Agriculture 44., pp. 189-202.

[13] Fang, C., Zhao-yan, L. & Yi-bin, Y., 2005, Machine Vision Analysis of Characteristics and Image Information Base Construction for Hybrid Rice Seed. In Rice Science., Vol. 12. Issue 1. pp. 13-18., China.

[14] Fei, D., Hengnian, Q. & Guangwu., Z.. 2012, Seed vigor assessment for Cunninghamia lanceolate and Pinus massoniana using image processing. In Transactions of the Chinese Society of Agricultural Engineering., Vol. 28. Supp. 2. pp. 274-279.

[15] Food and Agriculture Organization of the United Nations [homepage on the Internet]. Italy: Associate Online Resources. Retrieved November 17, 2013 available from: www.fao.org

[16] Geneve, R. L. & Kester, S. T., 2000, Evaluation of Seedling Size Following Germination Using Computer-aided Analysis of Digital Images from a Flat-bed Scanner. In HortScience. Unviversity of Kentucky. College of Agriculture journal., Vol. 36. Issue 6. pp. 1117-1120., Nov 2, 2000.

[17] Gomes Junior, F. G., Chamma, H. M. C. P. & Cicero, S. M.. 2014, Automated image analysis of seedling for vigor evaluation of common bean seeds. In Acta Sciencetiarum. Agronomy., ISSN: 1807-8621., Vol. 36. No. 2. pp. 195-200., Brasil.

[18] Gonzalez, R. C. & Woods, R. E., 2008, Digital Image Processing, 3rd Edition, ISBN: 9780131687288, Prentice Hall Publishing.

[19] Guzman, J. D. & Peralta, E. K., 2008, Classification of Philippine Rice Grains Using Machine Vision and Artificial Neural Networks. In World Conference on Agricultural Information and IT, IAALD AFITA WCCA 2008., Tokyo University of Agriculture, Tokyo, Japan.

[20] Hoffmaster, A. L., Fujimura, K., McDonald, M. B. & Bennett, M. A., 2002, An Automated System For Vigor Testing Three-Day-Old Soybean Seedlings. In MS. Thesis. The Ohio State University., Columbus, OH.

[21] Howarth, M. S. & Stanwood, P. C., 1993, Measurement of Seedling Growth Rate By Machine Vision. In Proceeding SPIE 1836. Optics in Agriculture and Forestry., Vol. 36 Issue 3. pp. 959-963.

[22] International Seed Testing Association [homepage on the Internet]. Retrieved November 8, 2012 available from: http://www.seedtest.org/en/home.html

[23] Justice, O. L., 1972, Essentials of seed testing. In Kozlowksi, T.T. (Ed.) Seed Biology., Vol. 3. pp. 301-370., New York, United States.

[24] Kaur, H. & Singh, B., 2013, Classification and Grading Rice Using Multi-Class SVM. In International Journal of Scientific and Research Publications., ISSN 2250-3153., Vol. 3. Issue 4.

[25] Kohavi, R. & John, G. H., 1997, Wrappers for feature subset selection. In Artificial Intelligence 97 Journal. ELSEVIER Science., pp. 273-324.

[26] Lilhare, S. F. & Bawane, N. G., 2012, Classification of Paddy Varieties using Image Processing. In National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2012), Proceedings published by International Journal of Computer Applications® (IJCA)., USA.

[27] Maheshwari, C. V. & Jain, K. R., 2013, Parametric quality analysis of Indian Ponia Oryza Sativa ssp Indica (rice). In International Journal for Scientific Research & Development (IJSRD)., ISSN: 2321-0613., Vol. 1. Issue 2.

[28] MathWorks: Accelerating the pace of engineering and science [homepage on the Internet]. United States: Associate Online Resources. Retrieved October 5, 2014 available from: http://www.mathworks.com

[29] Mladenov, M. & Dejanov, M., 2008, Application of Neural Networks for Seed Germination Assessment. In 9th WSEAS International Conference on Neural Network (NN'08)., Sofia, Bulgaria.

[30] Mo, C., Kim, G., Lee, K., Kim, M. S., Cho, B., Lim, J. & Kang, S., 2014, Non-Destructive Quality Evaluation of Pepper (Capsicum annuum L.) Seeds Using LED-induced Hyperspectral Reflectance Imaging. In Sensors. Open access journal., vol. 14. pp. 7489-7504.

[31] NPG: Nature Publishing Group [homepage of the Internet]. United States: Associate Online Resources. Retrieved April 29, 2015 available from: http://www.nature.com/subjects/biological-techniques

[32] Oakley, K., Kester, S. T. & Geneve, R. L., 2004, Computer-aided digital image analysis of seedling size

and growth rate for assessing seed vigour. In Impatiens. In Seed Science and Technology., Vol. 32. Number 3. pp. 907-915.

[33] Office of Agricultural Economics [homepage on the Internet]. Thailand: Associate Online Resources. Retrieved October 5, 2014 available from: http://www.oae.go.th

[34] OuYang, A., Gao, R., Liu, Y., Sun, X., Pan, Y. & Dong, X., 2010, An Automatic Method for Identifying Different Variety of Rice Seeds Using Machine Vision Technology. In Proceeding of the Sixth International Conference on Natural Computation., Vol. 1. pp. 84-88., Yantai, Shandong, China.

[35] Pandey, N., Krishna, S. & Sharma, S., 2013, Automatic Seed Classification by Shape and Color Features. In International Journal of Computer Applications Technology and Research., Vol. 2. Issue 2. pp. 208-213.

[36] Patil, N. K. & Yadahalli, R. M., 2012, Classification of food grains using HSI color model by combining color and texture information without performing pre-processing and segmentation. In World Journal of Science and Technology 2012., ISSN: 2231-2587., pp. 50-53.

[37] Punthimast, P., Auttawaitkul, Y., Chiracharit, W. & Chamnongthai, K., 2012, Non-destructive Identification of Unmilled Rice Using Digital Image Analysis. In the Proceeding of the 9th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology., pp. 1-4., Phetchaburi, Thailand.

[38] Rahman, M. M., Rahman, M. M. & Hossain, M. M., 2013, Effect of Sowing Date on Germination and Vigor of Soybean (Glycine max (L.) Merr) Seeds. In A Scientific Journal of Krishi Foundation. Index Journal., ISSN: 1729-5211., pp. 67-75.

[39] Rice Department [homepage on the Internet]. Thailand: Associate Online Resources. Retrieved November 17, 2013 available from: www.ricethailand.go.th

[40] Ricepedia - The online authority on rice [homepage on the Internet]. Retrieved November 17, 2013 available from: http://ricepedia.org

[41] Sansomboonsuk, S. & Afzulpurkar, N., 2006, The Appropriate Algorithms of Image analysis for Rice Kernel Quality Evaluation. In The 20th conference of Mechanical Engineering Network of Thailand., Nakhon Ratchasima, Thailand.

[42] Shantaiya, S. & Ansari, U., 2010, Identification Of Food Grains And Its Quality Using Pattern Classification. In International Journal of Computer & Communication Technology (IJCCT)., Vol. 2. Issue 2, 3, 4, 3-5, India.

[43] Silva, C. B., Lopes, M. M., Marcos-Filho, J. & Vieira, R. D., 2012, Automated system of seedling image analysis (SVIS) and electrical conductivity to assess sun hemp seed vigor. In Revista Brasilerra de Sementes., Vol. 34. Number 1. pp. 055-060., Londrina, Brazil.

[44] Silva, C. S. & Sonnadara, U., 2013, Classification of Rice Grains Using Neural Networks. In Proceeding of Technical Sessions, Institute of Physics., pp. 9-14., Sri Lanka.

[45] Silva, V. N. & Cicero, S. M., 2014, Imaging seedling analysis to evaluate tomato seed physiological potential. In Revista Ciencia Agronomica., Vol. 45. Number 2. pp. 327-334., Brasil.

[46] Thailand Project Encyclopedia for youth by the whim of the king [homepage on the Internet]. Thailand: Associate Online Resources [updated October 29, 1995]. Retrieved November 8, 2012 available from: http://kanchanapisek.or.th/kp6/

[47] The Post harvest Unit of the International Rice Research Institute (IRRI). Measuring seed germination - Rice Knowledge Bank [homepage on the Internet]. Associate Online Resources [updated August, 2011]. Retrieved from: September 12, 2013 available from: http://www.knowledgebank.irri.org

[48] Varma, V. S., Durgo, K. & Keshavalu, K., 2013, Seed image analysis: its applications in seed science research. In International Research Journal of Agricultural Sciences., Vol. 1. Issue 2. pp. 30-36.

[49] Yadav, B. K. & Jindal, V. K., 2001, Monitoring milling quality of rice by image analysis. In Computers and Electronics in Agriculture, ELSEVIER., Vol. 33. pp.19-33.

[50] Yao, Q., Chen, J., Guan, Z., Sun, C. & Zhu, Z., 2009, Inspection of rice appearance quality using machine vision. In Global Congress on Intelligent Systems, GCIS'09, IEEE Computer Society., Vol. 4. pp. 274-179., Xiamen, China.

[51] Zhao-yan, L., Fang, C., Yi-bin, Y. & Xiu-qin, R., 2005, Identification of rice seed varieties using neural network. In Journal of Zhejang University SCIENCE., ISSN: 1009-3095., pp. 1095-1100.

9. AUTHOR BIOGRAPHIES

Benjamaporn Lurstwut received her B.S. in ICT and M.S. in computer science from Mahidol University, Bangkok, Thailand. She is currently a Ph.D. candidate in computer science in the faculty of Information and Communication Technology, Mahidol University, Bangkok, Thailand. Her research interests include image processing, image segmentation, machine learning, and pattern recognition.

Chomtip Pornpanomchai received his B.S. in general science from Kasetsart University, M.S. in computer science from Chulalongkorn University and Ph.D. in computer science from Asian Institute of Technology. He is currently an assistant professor in the faculty of Information and Communication Technology, Mahidol University, Bangkok, Thailand. His research interests include artificial intelligence, pattern recognition and object-oriented systems.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 4, 183 - 190, 2016, ISSN:- 2319–8656

An Object Role Database Model for Enhanced Fuel Distribution and Sales Monitoring in Nigeria

Eke B. O.
Department of Computer Science
University of Port Harcourt
Port Harcourt, Nigeria

Egbono F.
Department of Computer Science
University of Port Harcourt,
Port Harcourt, Nigeria

Abstract: Central to any long – term stability in fuel scarcity is the effective production, distribution and monitoring of petroleum products sales across filling stations by nations with oil subsidy policies. The use of modern information and communication technology (ICT) acts as a facilitator in assisting the monitoring committee and the public in knowing which filling station has been allocated product and the quantity of petroleum product for any given filling station at any given time in the country. This paper, presents an Object Role Model of a highly scalable Distributed Database System designed to accommodate all data entities which include the quantity of fuel usage, number of filling station, quantity of fuel distributed, quantity imported, companies paid subsidy and companies owed. Other entities include quantity of petroleum refined locally and the location of the filling stations allocated. These data entities provide information that the public, fuel marketers, price monitoring team and the government can use in restoring normalcy in the chaotic fuel situation in Nigeria. The system will provide a central communication link among the stakeholders to enhance and facilitate harmony and the overall development of the petroleum sub sector of the national economy.

Keywords: Object Role Model, Distributed Database, Petroleum Distribution, Data entities

1. INTRODUCTION

An information system is an arrangement of people, data, processes, information presentation, and information technology that interact to support and improve day to day operations in a business as well as support the problem-solving and decision-making needs of management and users [1]. The definition highlighted a major point which is the support for problem-solving and decision-making needs of users.

In this paper, the users of the model are the developers of applications used by the public that consume the petroleum product and the government that need the information to stabilize the petroleum sector. The model supports a database entity technology that can be deployed in the development of information system that can aid several decision processes of fuel marketers and fuel buyers who engage in panic buying at any speculation of scarcity from rumors. Many organizations consider information systems and information technology to be essential to their ability to complete or gain a competitive advantage. Most businesses in their developmental effort and decision making process need to develop information systems by using the data entities in the process of the system development..

In building an information system, the stakeholders in the system need to be known. The stakeholders are the people who have interest in an existing or new information system. Stakeholders can be technical or non-technical and they can be broadly classified into six groups:

i) *System owners:* They pay for the system to be built and maintained. They own the system, set priorities for the system, and determine policies for its use. In some cases, system owners may also be system users.

ii) *System users:* They use the system to perform or support their work. They also define the business requirements and performance expectations for the system to be built.

iii)System designers: They design the system to meet the users' requirements. In many cases, these technical specialists may also be system analysts.

iv) System builders or programmers: They program, test, and deliver the system into operation.

2. ANALYSIS

The analyses involve the breakdown of conditions and the entities in such a way as to make them ready for use as components in the construction [2] of the fuel distribution and oil sales monitoring system.

2.1 Present System Analysis

Presently, Oil Marketers simply quote certain amount of money as their subsidy and the government just has to pay once it is certified by its agents, at least from the point of view of the public. The system could be very good if the government agents in-charge are professional and reliable. But in contrast the government agents could collide with the oil marketers to corruptly enrich themselves but over quoting the amount of money to be paid to get some cut from the marketers. However the system will add transparency to the process by linking the actual quantity of oil in circulation to the subsidy allowing all the stakeholders to estimate the actual subsidy that the government needs to pay.

2.2 Proposed System Analysis

There is a need for opening of the way to availability of fuel and the creation of free trade environment for the development of the petroleum industry in Nigeria. However within the subsidy regime there are data components that must be modeled to make the operation as open and efficient as possible via information technology.

The data components or entities includes [3]:

i) *The quantity of fuel usage*: If one asks any government agent the quantity of fuel that is actually used in the country, the answer will be a very bad estimation that will either be in gross excess or in greater underestimation. With this entity a specific figure need to be provided and the figure can be checked against other values from other sources.

ii) *Number of filling station*: When one moves around the nation there are many filling stations over grown with weeds showing that the filling station may exist only in NNPC papers but on reality does not exist. It is therefore extremely difficult to specify the actual number of filling stations operational in the country especially where some are temporarily or permanently sealed off by agents of government.

iii) *Quantity of fuel distributed*: Many filling stations that have fuel on paper are often closed and empty in reality and the fuel loaded for them to sell is diverted to other locations or have left the nation through illegal routes. The actual quantity of fuel in circulation is hardly known by the subsidy payer-government. But the entity offers a figure that can be checked against any other parameters.

iv) *Quantity Imported*: There are marketers who import product only in paper and as such collect certain amount of money from the government as subsidy for products that may have never existed in the first place. But if the quantity imported entity is specified then the depots that received the oil can easily be tracked to see if the fuel is actually on ground as shown by the system.

v) *Companies paid subsidy*: The companies that actually imported product must be the companies receiving subsidy and the money paid to them should be well specified so that it will be easy to know the exact outstanding fund so that the fire bridge approach of handling payment can be easily eliminated. Recently subsidy funds are released only when marketers go on strike or threaten to do so.

vi) *Companies owed:* The companies that are being owed need to be listed so that the companies will know that the government is aware of the debt and that they have intension to pay. The public will also expect the companies to either be patient or opt out of the subsidy system.

vii) *Quantity of petroleum refined locally*: The quantity of petroleum product needed to be imported can easily be computed if the quantity used are known and the quantity refined locally are known. The balance can easily be calculated and decision on how much is needed to be imported and pay for subsidy can be planned ahead and captured in the budget of the country for each year.

These data entities provides information that the public, fuel marketers, price monitoring team and the government can use in restoring normalcy in the chaotic fuel situation in Nigeria. The system will provide a central communicative link among the stakeholders to enhance and facilitate harmony and the overall development of the petroleum subsector of the national economy.

The entities will form the core of the database structure of the system to be developed for the petroleum subsector harmonization. The entities name may vary from one developer to another but the content may still remain the same. These entities are essential for the design stage of the database development required in the system development. It is also note worthy that the entities listed here are not necessarily the only entities required but are the major ones required.

3. DATABASE DESIGN

In the table 1 below the entities that are needed for the database are represented using their names, the entities types, there descriptions and the length.

The database record of the fuel activities listed in table1 is described below to provide further clarification which includes:

1) The serial number which can be used as a primary key for the database
2) The ability of identifying an individual fuel user is really a challenge and by extension the total number of fuel used in Nigeria. This is more challenging with illegal importation and illegal refining activities which add to the level of unaccounted fuel available in the nation. However ignoring theses sources and further blocking smugglers activities will provide a reasonable solution.
3) Number of filling station: The ability to record the functional filling stations and update the record when ever new licenses are granted for opening a new one is needed. Whenever a company opens another branch of their filling station the record must be updated. Since filling stations are physical it may be easier to get the data.
4) Quantity of fuel distributed: The filling stations needed to be modernized so that all fuel pumped can be recorded and automatically captured by the central system. This can be carried out by recommending modern electronic pump machines which can record and transfer the data to filling stations.
5) Quantity imported: The nature of import and discharge of fuel in the sea ports makes it very easy to record the quantity of fuel imported so that the trend of importation can be easy to record.
6) Companies paid subsidy: The Oil Companies that have been paid subsidy are listed with this field. The field

will help in knowing the budget required to pay subsidy ahead of time. The companies paid will also not have the opportunity to complain hiding on the shadow of the companies still owed.

7) Companies owed: The companies that have not being paid also needed to be listed under this field to offer this companies hope of payment and possibly the dates that payments are expected to be made. This helps the companies to sale on the normal price while waiting for payment.

8) Quantity of petroleum refined locally: The figures can be easily generated from all the refineries in the country. The figures are necessary to know the increase and decrease in refined product and the actual quantity needed to be imported.

3.1 Database Table Structure

In the design of the database there is a need to develop the table structure by extracting the possible fields expected from the database [4]. These fields are made of the names that are necessary for use in identifying the set of data needed to be presented in the database. The fields are supposed to keep data in the database with specific type which is usually specified in the entity type field. The field detail are also specified in the field description column and the length of each of the field is specified.

In table 1sample fields from different tables where illustrated using some of the important fields relating to the topic of discuss-oil sales monitoring in a subsidy country or regime. The table illustrate the field names, the entity types, the entity description and the length of the field. In real life DBMS the table description may vary slightly.

Table 1: Structure of Needed Database Files

SNO	Field/ Entity Names	Entity Types	Entity Description	Length
	SNo	Number	Serial Number of the Records	
1	**Qty_Fuel_Use**	Number	The quantity of fuel usage	
2	**Filling_Stat_No**	**Number**	Number of filling station:	
3	**Qty_Fuel_Distr**	**Number**	Quantity of fuel distributed:	
4	**Qty_Fuel_Imported**	**Number**	Quantity imported:	
5	**Companies_Paid**	**String**	Companies paid subsidy:	
6	**Companies_Owed**	**String**	Companies owed:	
7	**Qty_Refined**	**Number**	Quantity of petroleum refined locally:	

3.2 Entity-Relationship Conceptual Design

In other to depict the characteristics of database distributable over many sites, let us consider a particular case used for this research. We begin with a Government Fuel sales Monitoring/ Fuel Usage database. The database entities are illustrated in figure 1, the entities used are clearly listed and the relationship expected to bind the data is also illustrated. The entities include quantity of fuel used (Qty_fuel_used), Number of Filling station (Fillng_Stat_No) and with the input of Government, oil marketers and data gathering organizations. The databases interrelate with one another in various ways however this relationships can be represented as an entity-relationship diagram.

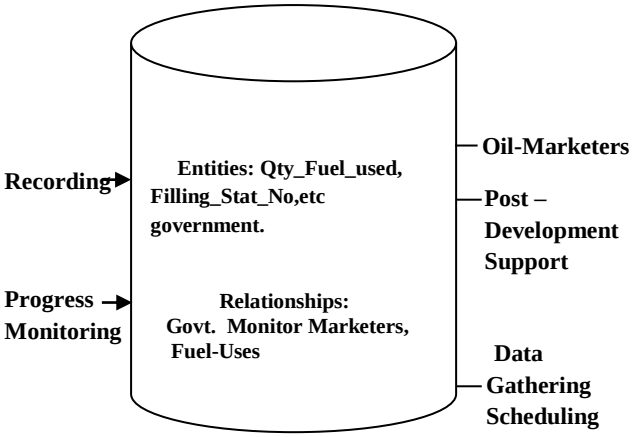


Fig 1: Depiction of an Fuel Monitoring Database

A fuel monitoring database for the system contains data about fuel usage, import distributions and other parameters that are already listed and required in keeping petroleum distribution and pricing stable across the nation. The Organizations that will gather the data, government, monitoring agents and the oil marketers are also captured in the system. The design presented however is the conceptual design and not the detail design of the system. The detail ERD design is presented in the next section. The implementation of the database may be at different locations or different Network servers but due to their relationship, many entities can be fetched, joined and used in generating answers to various kinds of questions which ordinarily would have been difficult to answer without the distributed database [5].

3.3 Design View and Integration

The distributed database is not generated at a single site, therefore it can be classified as large and complex [5]. Therefore, there must be a way to manage the complexity of the design process. Design View and integration can be used in managing the fuel monitoring and usage database design project by providing a way to break a large effort into smaller parts. In these cases, we combine individual views into a complete database design.

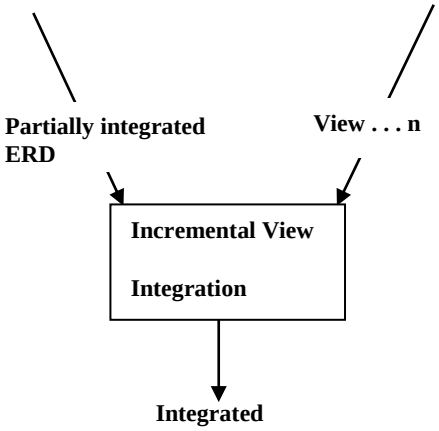


Fig 2: Incremental Integration Design

In Figure 2, a view and a partially integrated Entity Relationship Design (ERD) are merged in each Integration step. Initially, we choose a view and contracts on ERD for it. For instance, filling stations can comprise fuel consumers and oil Marketers data on fuel sales, fuel bought and fuel supplied ; these can be different views of the same database. Each view has a different ERD and is incremented as a partially integrated ERD produced after each step. The approach can be binary as the current view is analyzed along with the partially integrated ERD or multiple views. ERDs are produced for each view and then the view ERDs are merged. The integration occurs in one large step after all views are analyzed. This approach is parallel because different designers can perform view designs at the same time.

In this approach we postponed integration until the end when all views are integrated to produce the final ERD. The incremental approach [5] is well suited for the implementation in this work because of closely related views. For example, the fuel importation, fuel usage, filling station and subsidy payments are closely related, because importation precedes payment and availability in filling station which influences fuel usage.

4. Entity-Relationship Detail Design

In the detail design of the ERD design presented in figure 3 the Oil filling station is related to the Consumers table by sales of fuel. The relationship is defined as one or more oil filling stations relating to one or more fuel consumers. This relationship is clearly illustrated. On the other hand the Oil depot supplies oil to the oil filling stations and the relationship is clearly shown as non or more depot to oil filling station relationship. This indicate that all fuel supplied at filling stations do not necessarily originate from depots others may be direct shipment from refineries to the filling stations.

In figure 3, the relationship of oil monitors and oil filling station clearly shows a one to many relationship where one filling station get checked by one or many monitoring agents and many filling station also get checked by one or more monitoring agents using check compliance from the monitor. Here the rules and price limits for the sale of the product is being monitored to ensure that filling stations complied to the regulations specified. The relationship between importers and oil depots is also clear, importers ship their imported fuel to oil depots from where they are supplied to the oil filling stations. The oil depots in turn expect to get oil from the importers.

In Nigeria, importation of oil is handled by independent oil marketers and the nations oil company –Nigerian National Petroluem Company (NNPC). The percentage of importation usually vary from 50-50, 30-70 and even 0-100 depending on policy of the day and the availability of foreign exchange for importation of mostly refined product. Unlike many developed and developing nations, Nigeria exports crude and imports refined product since its locally refine product could not sustain local usage at least as at the time of this publication. However, if all the product are refined locally the situation may remain the same if the subsidy regime persists. The Department for Petroluem Resources (DPR) is saddled with the responsibility of monitoring the sales of petroleum products at government approved price.

In the entity relationship diagram it is equally clear that the relationship between the Independent (Ind) marketers and the importers of fuel are demonstrated. The independent marketers import oil and therefore are importers and share data and information with the importer database.

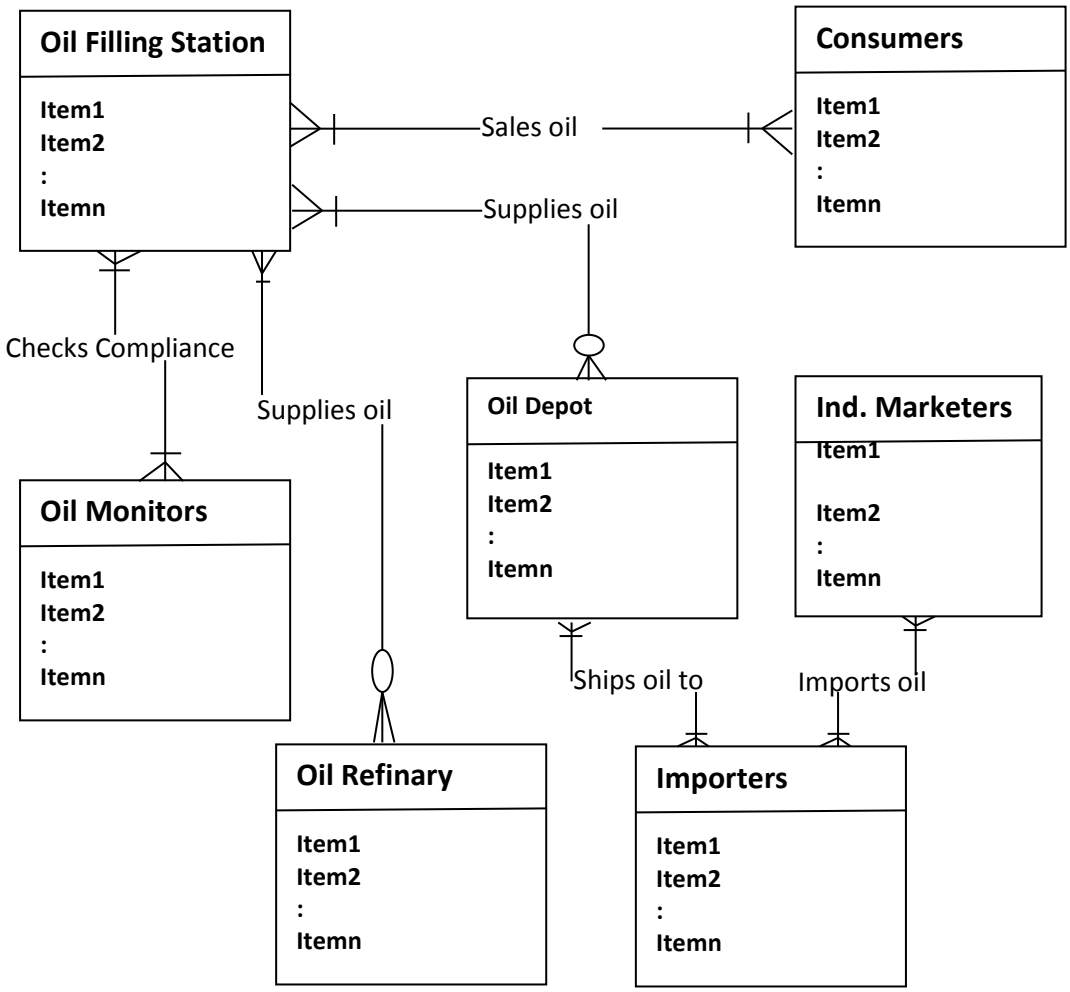


Fig. 3: The Entity Relationship Diagram of the Distributed Database of Oil Sales Monitoring System.

5. OBJECT ROLE MODEL DESIGN

Object Role Model (ORM) is a design system at the conceptual level and mapping between conceptual and logical (e.g. relational) levels, where the application is described in terms readily understood by users, rather than being recast in terms of implementation data structures. ORM includes a formal, textual specification language for both models and queries, as well as a formal, graphical modeling language [6].

Object Role Modeling got its name based on its views of the application world as a set of objects (entities or values) that plays roles (parts in relationships). It is sometimes called fact-based modeling because ORM verbalizes the relevant data as elementary facts. These facts cannot be split into smaller facts without losing information.

Object role modeling recognizes four basic kinds of data objects: simple, value, composite, and nested. A simple object is one in which real world instances are designated -- uniquely identified -- by a single data element; i.e., a single data element comprises the primary key [7]. Figure 4 shows how oil filling station would be represented in an object role model when real-world filling stations are to be designated in the database by an identifying number, Station_id, rather than by their names. Here, an oil Filling station is an example of a simple object.

The ORM model design of the fuel distribution and fuel sales monitoring system is presented in figure 4 using an Object Role Modeling technique. This design makes it easy to understand the movement of data activity and interaction between objects without presenting the detail data components of the objects.

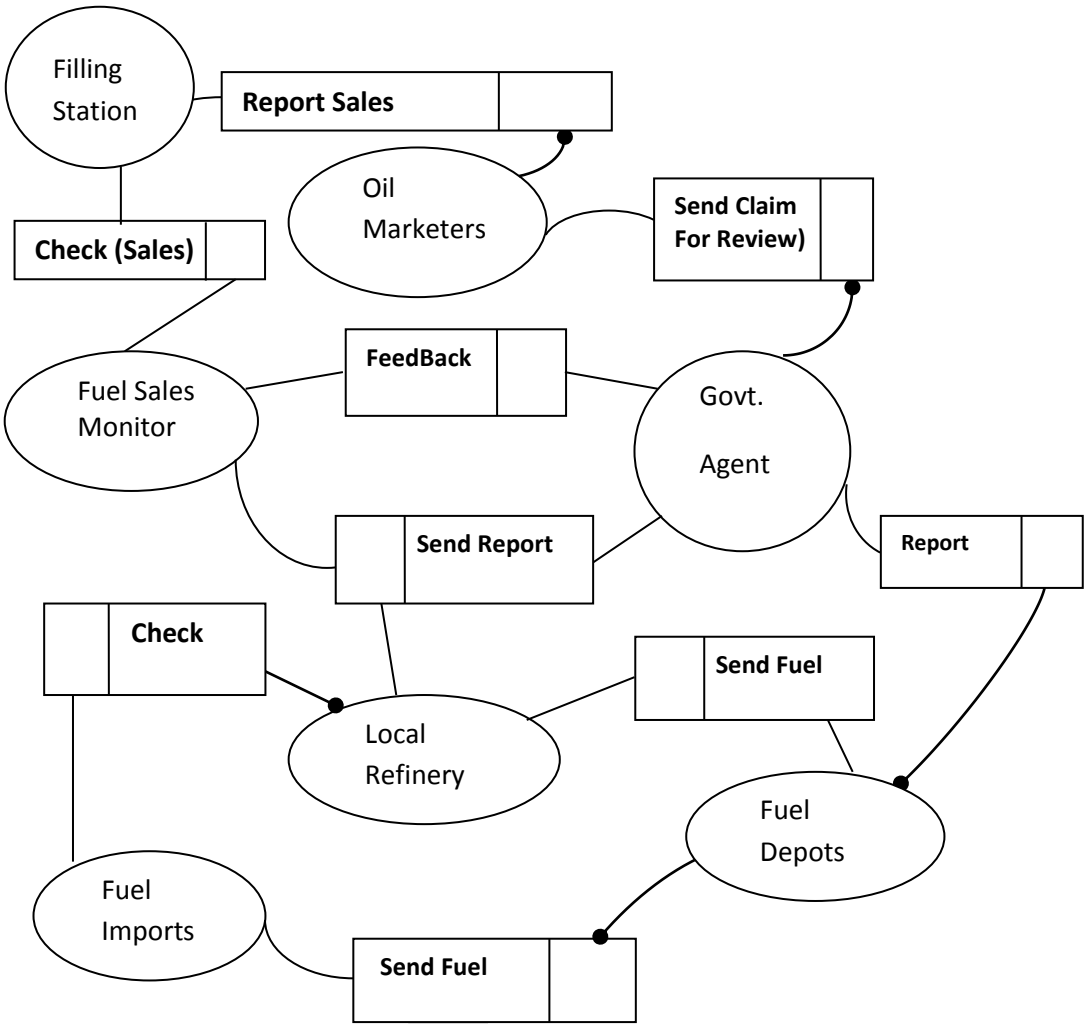


Fig 4: Design of the Object Role Database Model using ORM

A value object is something like a name (label), number, or date, which represents a simple scalar attribute [8]. Fuel Depots in Figure 4, is a value object. It is clear from the object-oriented perspective that an object- Fuel Imports is distinctive from Fuel Depots yet there is a relationship. It is shown from the model that Send Fuel links the two distinct objects since fuel imports (imported fuel) are sent to fuel depots. When simple and value objects are combined using a key, composites are formed. An object whose existence and identity stems from the relationship between two other objects is called a nested object. A "circle" (a rectangle with rounded corners) is drawn around the relationship to "objectify" it. A nested object can then participate in relationships with other objects [9].

In figure 4, several relationships exit between various objects within the object-role model. This relationships include report sales connecting the oil marketers and the filling stations, check sales connecting the fuel sales monitors and the filling stations, and send claim which links the oil marketers and the Government agent responsible for payment of subsidy claims. Other relationships include feedback connecting the fuel sales monitors (DPR in Nigeria) with the Government agent incharge, report linking the fuel depots and the Government agent, send fuel linking the depots and the local oil refinery and check connecting fuel imports and the local refinery. This relationships are easy to understand by both the developers and the requirement specifier or system owner.

The object-Role Model design presented in this work serves as an easy pedestal for the application developers or the programmers to leverage in the implementation of the system devoid of details of the each database. This makes the system adaptable to various implementation for a particular country based on the actual data available on fuel subsidy in the country. The model can also be applied in process to process assessment of the system before commitment can be made on implementation.

6. CONCLUSIONS

In Nigeria a lot of effort has been made to improve the fuel supply with mixed success. There is the need to develop a system that will guide the Government authorities, Independent Petroleum Marketers Association of Nigeria (IPMAN)- oil marketers, Department of Petroleum Resources (DPR)- fuel monitors and other stakeholders in the distribution of fuel in Nigeria. In the first instance, we need to design the system before its implementation. In this paper, the Object Role Model has been designed for the system for monitoring fuel sales by conceptualizing the system, producing the Entity relationship and using the entities in the development of the ORM for the system. The developed system can be implemented based on the policy of the government on fuel subsidy at any given time. In Nigeria, Government and Independent Marketers jointly import subsidized refined products at varying ratios, such as 50-50, 70-30 and even 100-0 for government against marketers. At 50-50 the government oil company –NNPC imports 50% of the product while marketers imports 50%. At 100-0, only the imported fuel of NNPC is subsidized, if the marketers import fuel their fuel imports are not subsidized at all. It shows from the design that the implementation is expected to follow the policy being carried out at a given time.

6.1 Recommendation

In this paper, the users of the model are the developers who implements the system for fuel sales monitoring either from the government or from the independent marketers. The government need the implemented system to stabilize the petroleum distribution challenge in Nigeria. The model will offer a database entity technology that can be deployed in the development of information system that can aid several decision processes of fuel marketers and fuel buyers who engage in panic buying at any speculation of scarcity from rumors. The government can also use the system to reduce subsidy misunderstanding and mis-information between government, the public and oil-marketers.

Model designers who want to develop other systems in information technology using Object Role Modeling technique will find the work in this paper useful, since it is simple and direct to the application area. The designers can derive inspiration from the work in analyzing their systems and from the entities of that particular system design the ORM model which can be useful in the development of their systems.

The model designed in this paper can be easily modified for application in other related areas involving government subsidies [11] such as Bread Subsidy (in Egypt), feeding subsidy (in India), agricultural subsidy (in USA) and other government subsidies in other countries.

7. ACKNOWLEDGMENTS

Thanks to Oyol Computer Consult, Inc Port Harcourt, Nigeria for supporting the research in the time when power and petroleum scarcity seem to be most critical in Nigeria. We also expect their contribution in the implementation of the model designed in this research.

8. REFERENCES

[1] Forouzan B. A. and Fegans, S. C (2005) Data Communicational and Networking, Tata McGraw-Hill Publishers Companies Ltd New Delhi 11008, (2005). 8-9, (Third Edition), Website: www.tatamcgrawhill.com.

[2] Pratt, Philip J., and Adamski, Joseph J.,(1991) Database Systems - Management and Design, 2nd Edition, Boyd and Fraser, 1991.

[3] Adeniyi G. (1997): "The Impact of Oil Exploration and Production Activities on the Environment: Implications for Peasant Agriculture" Seminar Paper on Oil and The Environment organized by Friedrich Ebert Foundations in Port Harcourt.

[4] Kossmann D., (2000).The State of Art in Distributed Query Processing. ACM Computing Surveys. Volume 32, No 4 December pp 422-469

[5] Gupta, A. (2009). Database Management System in the Practical Approach to SQL & PL/SQL. Daryaganj Delhi: S. K. Kataria & Sons.

[6] Terry Halpin (2012) Object-Role Modeling (ORM/NIAM), Microsoft Corporation, USA reproduced by permission.from Handbook on Architectures of Information Systems, eds P. Bernus, K. Mertins & G. Schmidt, Springer-Verlag, Berlin, 1998, www.springer.de/cgi-bin/search_book.pl?isbn=3-540-64453-9.

[7] Stanley D. B. (2015) A Primer on Object Role Modeling, University of California Press,Berkeley, USA, accessed 2015

[8] Garvey, M. A., and Jackson, M. S. (2010), " Object-Oriented Databases", Information and Software Technology, Vol. 31, No. 10, pp524-525 Dec. 2010.

[9] Akash M. (2012)Classifying data for successful modeling, http://www.dwbiconcepts.com/data-warehousing/12-data-modelling/101-classifying-data-for-successful-modeling.html

[10] Gourevitch, P. 1986. Politics in Hard Times : Comparative Responses to International Economic Crises. Ithaca: Cornell University Press.

[11] Beñat B., Soumitra D., and Bruno L. 2013 The Global Information Technology Report 2013, Growth and Jobs in a Hyperconnected World, Pulished by World Economic.

9. ABOUT THE AUTHORS

Dr. Eke Bartholomew is a Software Engineering Lecturer in Uniport and visiting Instructor at Institute of Petroluem Studies UPH. His research interest is in SE Methodologies and System Devepolment. He has numerous publications in Nigeria and across the Globe.

Dr. Fubara Egbono is a Lecturer at Department of Computer Science, University of Port Harcourt. He specializes on Distributed Databases and Machine Architecture. His research interest is in Solving real life challenges using modern data design principles.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 4, 191 - 197, 2016, ISSN:- 2319–8656

Classification of Name Based On Leaf Recognition Using BT and ED Algorithm

A.M.Ravishankkar
Karpagam University,
Coimbatore,Tamilnadu
India

M.Mohanapriya,
Karpagam University,
Coimbatore, Tamilnadu,
India

ABSTRACT:

The main purpose of this paper should be to show that the outer frame of a leaf and with the help of Back propagation Network is enough to give a reasonable statement about the species category is identified. Leaves Recognition is a neuronal network based java application/applet to recognize images of leaves using Back propagation Network. The intention is to give the user the ability to administrate a hierarchical list of images, where they can perform some sort of image using edge detection to identify the individual tokens of every image. The Thinning algorithm here is used to process the image recursively and minimizes the found lines to a one-pixel wide one by comparing the actual pixel situation with specific patterns can be identified and then minimizes it. The urgent situation is that due to environmental degradation and lack of awareness, many rare plant species are at the risk of extinction so it is necessary to keep record for plant protection. It focuses on using digital image processing for the purpose of automate classification and recognition of plants based on the images of the leaves. It help to protect the plant and mainly it is used for highly production of rare plant or herbal plant used for medical purpose. Efficacy of the proposed methods is studied by using two neural classifiers. These are neuro-fuzzy controller and a feed-forward back-propagation multi-layered perception to discriminate between 28 classes of leaves. The features have been applied individually as well as in combination to investigate how recognition accuracies can be improved with the help of B&T algorithm.

Keywords: Image Processing, Edge detection, Neuronal network ,thinning algorithm, B&T algorithm.

1. INTRODUCTION:

Plants play the most important part in the life cycle of nature. They are the primary producers that sustain all other life forms including animal, people and also Non living things. This is because only plants are the only organisms or species that can convert light energy from the sun into food. Both human and animals are incapable of making their own food, depend directly or indirectly on plants for their supply of food. Leaves of same species also have variation in there shapes and moreover leaves of different species may have a same size because of the complex nature of leaves. So it is very difficult to identify plant name for that we need higher process of computing the leaves with efficient technique. A leaf from an unknown species of plant will be the input to the proposed system and trained set is to identify plant recognition scheme based on the trained system.To handle such volumes of information ,development of a quick and efficient classification and decision based method has become an area of active research. In nature, plant leaves are two dimensional containing important features that can be useful for classification of various plant species, such as shapes, colours, textures and structures of their leaf can be varied from one to another [1,4]. A leaf from an unknown species of plant will be the input to the proposed system. The system then segments the leaf image from its background, computes the

morphological feature representation help for matching the leaf, and then displays the similarity percentage as computed [5]. The leaf image will be captured on a plain contrast background to reduce the complexity of the segmentation algorithm and give better performance.

The present paper proposes a scheme for automated recognition of three types of plant species by analyzing shape features from digital images of their leaves with the help of enhanced algorithm. It follows: Section 2 provides an overview of related work, Section 3 outlines the proposed approach with discussions on overview, feature computation and classification schemes, Section 4 provides details of the dataset and experimental results obtained and Section 5 provides the overall conclusion and the scope for future research.

2. RELATED WORK:

Beside the fact of writing a java based application to realize this purpose, one additional feature is that it could also be used as a java applet to directly give the user the ability to start it via a java enabled internet browser. The main tasks of this application is used to detect the tokens using prewitt edge detection algorithm[2].A plant leaf identification, most of them used neuronal network algorithm. Image processing is most important preliminary phase and it taking image as a tokens.. This tokens will then be the basis of the neuronal network calculations to make it possible to recognize a unknown leaf image and specify the species it belongs to.This paper implements a leaf recognition algorithm using easy-to-extract features and high efficient recognition algorithm. Our main improvements are on feature extraction and the classifier. All features are extracted from digital leaf image. Except one feature, all features can be extracted automatically[7]. Jyotismita Chaki, Ranjan Parekh [9].In this paper leaf recognition has been done by using shape analysis and feature extraction. With the help of netural network which under supervised learning algorithm having multilayer preceptor weight can be verified with feed forward backward back propagation architecture.

Most popular algorithm to identify name for the plant with the help of leaf. General regression neural networks perform regression where the target Variable of the two values is continuous. The main aim is to predicted target value of an item(leaf) is similar to be the same as other items(leaf) that have close values of the predictor variables. The k-nearest neighbor algorithm is the most popular algorithms is to identify object (leaf) which is common to all of the other leaf have a similar properties to match with another. Gabor wavelet/AN Networks system to classify images texture to identify the name of the plant . K-means clustering is an algorithm to classify the objects based on attributes/features into *K* number of groups where *K* is a positive integer. K- means clustering is a supervised learning algorithm and it have a prior knowledge of the number of clusters maximizing intra clustering and Minimizing inter clustering.
In the neuronal network is used for sigmoid function[12].Plant species identification requires recognizing the plant by various characteristics, such as size, form, leaf shape, flower color, odor, etc., and linking it with a common or so-called scientific name[8]. The classification algorithm implemented for accurate identification of the plants based on Leaf image. Different data modelling techniques used include curvature scale space, fuzzy logic, fractal dimensions[9], Fourier analysis ,wavelets [9,11], curvelets and Zernike moments . A variety of classifiers have also been used viz. neural networks [3], support vector machines , nearest neighbours [6], and K-means for identifying unknown leaves[10].

3. PROPOSED SYSTEM:

Our system is based on image processing which finds an unknown leaf species without any previous knowledge, which is useful for any layman. The basic factors for identification of species are, image edge detection , back propagation and Neural Network shown in Fig 1.It is used to retrieval of leaf images based on the shape of the leaf image given as input by the user. For example, If the input is a unknown leaf‘s image, then the output will be given that determine the which leaf is present here.

3.1 Converting RGB image to binary image

The leaf image is acquired by scanners or digital cameras. Since we have not found any digitizing device to save the image in a lossless compression format, the image format here is JPEG. All leaf images are in 800 x 600 resolution better for identify for the purpose. There is no restriction on the direction of leaves when photoing.

In this approach we follows several steps,

1. user uploads an image of leaf of his /her choice.
2.we find out the feature extraction using image preprocessing.
3. we find out the image edge detection and thinning algorithm.
4.comparing the training data set with test data set.
5.Then determine the information about the particular image.

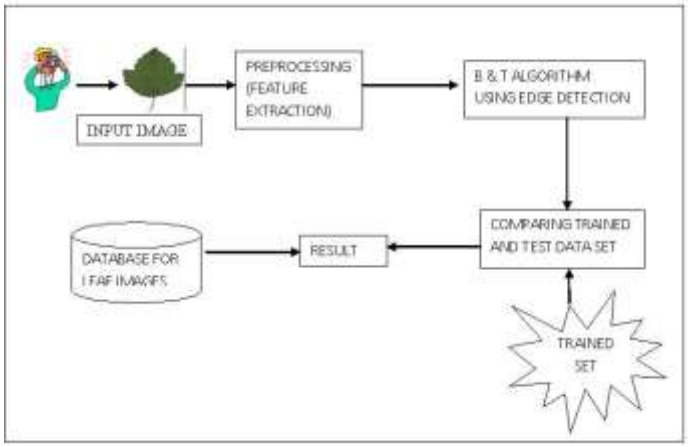


Fig 1.Architecture Diagram

STEP 1: Image Upload

User can upload any image of a leaf in jpeg or png format of his or her choice. We have ready dataset for comparing with that unknown image.

STEP 2: Image Preprocessing

Image processing is the enhancement of image i.e., processing an image so that the results are suitable for particular application. Processing image involves operations for making the background white for removing unwanted noise.One of the most important parts of the whole application is the image processing. Without finding any usefull tokens in the leaf images, the neuronal network calculation. So we spent lots of our efforts in the edge detection and thinning algorithms.

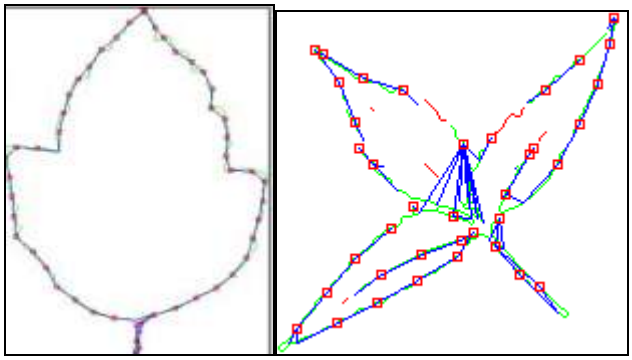


Fig 2 Image edge detection

In edge detection, the distance of the tokens (red square) and a minimum amount of pixels a line have to be to be recognized as a part of the shape. For a good quality result you should normally add at least 5 images of a species to give the neuronal network enough tokens to find the specific shape of this leaf species.

STEP 3: Edge Detection

The proposed system finds the specific tokens which represents that portion of the image where transfer of lower to higher pixel intensity occurs. These tokens are used for neural network calculations. Edges are recognized by abrupt intensity variations in an image. The first gradient is typically calculated to locate the edges. The 3 * 3 masks are used to extract the edges from image and convolved them through it. Prewitt edge detection technique is selected due to efficiency and simplicity in single template.

Prewitt edge detection produces an image (Fig2) where higher grey-level values indicate the presence of an edge between two objects. The Prewitt Edge Detection filter computes the root mean square of two 3x3 templates. It is one of the most popular 3x3 edge detection filters.
The Prewitt edge detection filter uses these two 3x3 templates to calculate the gradient value:

-1 0 1 1 1 1

```
-1  0  1      0  0  0
-1  0  1     -1 -1 -1

   X           Y
```

Now consider the following 3x3 image window:

```
+------------+
| a1  a2  a3 |
| a4  a5  a6 |
| a7  a8  a9 |
+------------+
```

where:

- a1 .. a9 - are the grey levels of each pixel in the filter window
- X = -1*a1 + 1*a3 - 1*a4 + 1*a6 - 1*a7 + 1*a9
- Y = 1*a1 + 1*a2 + 1*a3 - 1*a7 - 1*a8 - 1*a9
- Prewitt gradient = SQRT(X*X + Y*Y)

All pixels are filtered. In order to filter pixels located near the edge of an image, edge pixels values are replicated to give sufficient data. The idea behind the transfer of the leaf image shape into a neuronal network usable form is, that the cosines and sinus angles of the shape represents the criteria of a recognition pattern.

The right hand image shows a part of a leaf image that was already processed through the above mentioned edge detection and thinning algorithms.

To give you an idea of what you see in this image,

- Green line: The shape of the leaf image after successful edge detection & thinning.
- Red Square: This square represents a point on the shape of the leaf image from which we are going to draw a line to the next square.
- Blue line: The compound of the center of two squares from which we are going to calculate the cosines and sinus angle. Such a blue line is a representation of a leaf token.

Mainly this configuration is the properties of the neuronal network. It based on the amount of images and network properties you normally need to specify around 500-1000 training steps to get a good result in the recognition later. If the error rate drops below 0.01 you normally should encounter no problem in recognizing different leaf images.

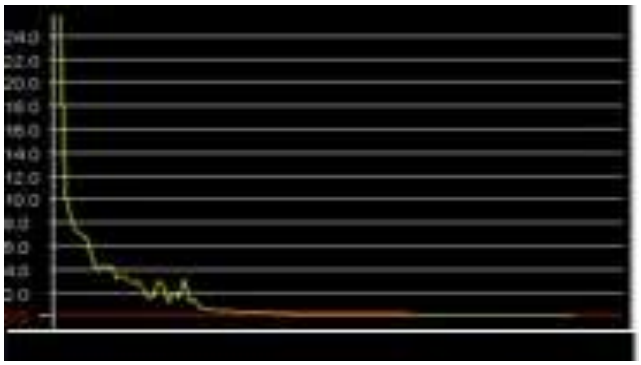


Fig 3 Detection graph

STEP 4: Neuronal network

Another main part of this work is the integration of a feed-forward back propagation neuronal network. As described earlier the inputs for this neuronal network are the individual tokens of a leaf image, and as a token normally consists of a cosinus and sinus angle, the amount of input layers for this network are the amount of tokens multiplied by two. The number of output neurons is normally specified by the amount of different species because we use a encoded form to specify the outputs. All other behaviour of the network is specified by the normal mathematical principals of a back propagation network. This neuronal network adopts with three tier network structure including the input layer, hidden layer[7],output layer.

In our system we are using neural network with one hidden layer. Each hidden layer is associated with the sigmoid function. In other words neurons in a same layer have same activation function. Sigmoid function is an exponential function which is used for calculation and transfer of knowledge from input neurons to output neurons. The graph for sigmoid function can be shown as in Fig 3.

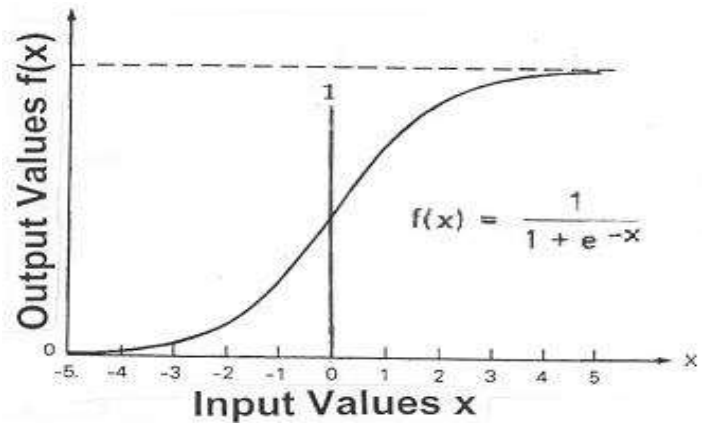


Fig 3 Sigmoid function

Sigmoid Function: f (x) = 1/(1+e^-x)

Sigmoid Derivative: f'(x) = f (x) * [1 - f (x)]

STEP 5: Thinning

To identifying a specific leaf image`s species here is that the outer frame of a leaf is enough to specify the species it belongs to. To accomplish that, it is necessary to identify this outer frame exactly. The previously applied Prewitt Edge detection normally just identifies the edges with a preconfigured threshold and after this edge detection we have to perform a thinning algorithm to minimize this threshold-based edge to a one-line frame.

The used thinning algorithm here processed the image recursively and minimizes the found lines to a one-pixel wide one by comparing the actual pixel situation with specific patterns and then minimizes it.

4. EXPERIMENTAL RESULT:

We have used the dataset for various leaf species like Azadirachta indica(neem), Pinus(pine tree), Quercus(oak), etc . Here, Firstly we add new species then add images of the same species under it and find the tokens for each leaf image after that we proceed by training these tokens using neural network later in recognition panel we add the unknown leaf to be recognized.

In order to optimize obtained results, we used to combine these features, where we get more efficiency in classification; the following table and figure prove this idea.

Algorithms	Accuracy	Precision	Recall	Fmeasure
naive bays	85.21	85.45	83.87	81.67
decision tree	68.88	56.09	87.45	45.78
prewitt edge detection	88.34	78.9	79.23	82.32
back propagation /thinning	90.45	92.78	91.1	92.45

Table 1 Results obtained by classification of edge detection algorithm

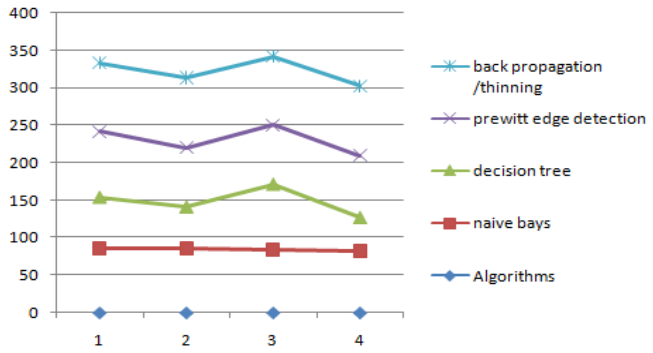


Fig 4 Results obtained by classification of edge detection algorithm

We compared the accuracy of our algorithm with other general purpose (not only applicable to certain species) classification algorithms that only use leaf-shape information. According to Table 1, the accuracy of our algorithm is very similar to other schemes. Considering our advantage respect to other automated/semi-automated general purpose schemes, easy-to- implement framework and fast speed of B & T algorithm, the performance is very good.

5. CONCLUSION:

Plants play an important role in our lives, without plants there will not be the existence of the ecology of the earth. The large amount of leaf types now makes the human being in a front of some problems in the specification of the use of plants, the first need to know the use of a plant is the identification of the plant leaf.The above analysis and graph, we display various details of unknown species in a specified area. For future scope diseases occurring in the unknown species and solution to overcome the diseases will provided and high production of rare plant. Back Propagation and Thinning algorithm must be done which gives output in terms of very high accuracy using minimal computational resources.

REFERENCES

[1] Jyotismita Chaki, Ranjan Parekh.:'Plant Leaf Recognition using Shape based Features and Network classifiers',(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 10, 2011.

[2] Stephen Gang Wu, Forrest Sheng Bao, Eric You Xu, Yu- Xuan Wang, Yi-Fan Chang and Qiao-Liang Xiang.:'A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network',arXiv:0707.4289v1 [cs.AI] 29 Jul 2007.

[3] J. Chaki, and R. Parekh.:'Plant leaf recognit ion using shape based features and neural network classifiers', Internat ional Journal of Advanced Computer Science and Applications (IJACSA), 2011, pp.26-29.

[4] Sandeep Kumar's.:'Leaf color, area and edge Features based approach for Identification of Indian Medicinal plants',IJCSE Vol 3 No.3, 2012.

[5] S. Beucher, F. Meyer.:'The morphological approach to segmentation: the watershed transformation', E.R. Dougherty (Ed.), Mathematical Morphology in Image Processing, Marcel Dekker, 1993

[6] X.F. Wang, J.X. Du, G.J. Zhang.:' Recognition of leaf images based on shape features using a hypersphere classifier', Computer Sci. 3644(2005) 87–96.

[7] Mokhtarian, F., & Abbasi, S.:'Matching shapes with self-intersections:Application to leaf classification', IEEE Transactions on Image Processing, 13(5), 653–661,2004.

[8] Lee, C.-L., & Chen, S.-Y.:'Classification of leaf images', International Journal of Imaging Systems and Technology, 16(1), 15–23 (2006).

[9] X. Gu et al.,Wang.:'Leaf recognition based on the combination of wavelet transform and Gaussian interpolation', *ICIS*, vol. 3644/2005, pp. 253-262, 2005.

[10] Biva shrestha" classification of plants using images of their leaves" IJCSITS, Vol. 2 No. 2, 2012

[11] M. T. Hagan, H. B. Demut, and M. H. Beale, Neural Network Design, 2002

[12] Casanova, D., de Mesquita Sá Junior, J. J., Bruno, O. M.:'Plant leaf identification using gabor wavelets', International Journal of Imaging Systems and Technology, 19(3), 236–243.

[13] Sigmoid Function: http://www.molecularcancer.com

[14] H. QI and J.-G. YANG.:'Sawtooth feature extraction of leaf edge based on support vector machine', Second International Conference on Machine Learning and Cybernetics, November 2003.

[15] S. M. Hong, B. Simpson, and G. V. G. Baranoski.:'Interactive venation- based leaf shape modeling', Computer Animation and Virtual Worlds, vol. 16, 2005.

[16] F. Gouveia, V. Filipe, M. Reis, C. Couto, and J. Bulas-Cruz.:'Biometry: the characterisation of chestnut-tree leaves using computer vision', IEEE International Symposium on Industrial Electronics, Guimara˜es, Portugal, 1997.

[17] X. Gu, J.-X. Du, and X.-F. Wang.:'Leaf recognition based on the combi- nation of wavelet transform and gaussian interpolation', International

Conference on Intelligent Computing 2005, ser. LNCS 3644. Springer, 2005.

[18] D. F. Specht.:'Probabilistic neural networks', Neural Networks, vol. 3, 1990.

[19] R. C. Gonzalez, R. E. Woods, and S. L. Eddins.:'Digital Image Processing Using MATLAB', Prentice Hall, 2004.

[20] T. Master.:'Practical Neural Network Recipes'. New York: John Wiley, 1993.

.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 4, 198 - 201, 2016, ISSN:- 2319–8656

Automation Testing In Software Organization

Prasad Mahajan
Bharati vidyapeeth COE
Pune, India

Harshal Shedge
Bharati vidyapeeth COE
Pune, India

Uday Patkar
Bharati vidyapeeth COE
Pune, India

Abstract: In software testing, automation testing plays a great role for improving test efficiency of the software testing team. Sometimes manual testing may not be effective due to its inconsistency, lack of coverage and none repeating in nature. To overcome this Test automation is used in software industry. In this paper we will discuss about Test Automation its pre-requisites, working steps, when to use automation testing, benefits over manual testing and selection of test cases to automate Thus *there are a number of testing tools available in the market out of which we will also discuss about* Selenium automation tool

Keywords: Test automation, Testing tools, Automation testing, Test case, Manual testing, Pre-requisites, Selenium

1. INTRODUCTION

Software testing is a set of activities conducted for finding errors in software. It is a process used to measure the quality of the software. Manual testing and automation testing are the two ways of testing. Manual testing is also called as static testing. It is carried out by the tester. Automation testing is also called as dynamic testing. But the problem is it is very time consuming process and requires more effort. So, automation testing is used to solve these problems. Automated testing is divided into four types such as reliability testing, security testing, correctness testing, and performance testing. It automates the steps of manual testing using automation tools . Automated tests are fast to execute and they are repeatable in nature. There are various tools available in the market which are used to test the process and targeted specific test environment. The environment may be functional, performance or exceptional testing etc. Testing tool should be selected on the basis of its compatibility with checklist for that purpose pilot round of the corresponding tool should be done. Cost is also an important factor for selection of tool.

2. AUTOMATION TESTING

2.1 Test Automation stages[3]

Please Normally the Automation manager will be responsible for selection of tool, Test engineer will be responsible for script generation, deployment and execution this stage will decide whether we need to automate the project or not. If failure occurs in this stage then it creates a larger impact on the project execution.

Following figure shows various stages of a test automation process.

Figure1. Stages of Automation

2.2 When to Automate [3]

1. After all manual testing process is in place as per expected results

2. At least tested twice

3. Environmental/Build should be constant or stable

2.3 Benefits over manual testing [2]

Following are some of the benefits of test automation

1. Fast application development by reducing time for testing

2. Repeatable in nature

3. Reliable

4. Scripts are reusable

5. Programmable

6. Provides high coverage for regression testing without adding additional resources

7. Improvement in productivity

8. Detailed test logs

9. Execution of the scripts across multiple platforms

10. Cost effectiveness improvement

2.4 Test cases to automate [3]

Selection of correct test cases for automation is very important as it impacts on development cost and scheduled time. One time testing is not done by automation tool. Also automation testing is not use for Usability testing to check "How easy is the web site to use?". This testing technique is not useful for tests which don't have predictable results.

Tests that need to be run for every build should be selected for automation.

Test that uses multiple data values for the same action are also needed for automation.

Similar tests that need to be executed using different platforms.

Regression testing is done by automation tool.

2.5 Applications [1]

Automation testing is use for testing web based application for finding flaws in the application and also to check the security of the system.

It provides a complimentary approach to manual testing to use automation testing to improve the level of test automation and reduction of risk instead of using more people.

For test management automation testing provides different types of tools.

Effective defect tracking is done by automation tool.

Table1. Automation Testing Vs Manual Testing

Information	*Automation testing*	*Manual testing*
Testing types	Regression testing	Usability testing
Execution speed	Fast to execute	Slower than automation testing
Sequence	After manual testing	Before automation testing
Resources	Tool should be used instead of large number of human resource	Large number of human resource is required

3. SOFTWARE TESTING TOOLS [2]

3.1 Selenium

Selenium is an open source tool.

It is a robust set of tools that supports rapid development of test automation for web-based applications. Selenium Supports Cross Browser Testing as well as selenium tests can be run on multiple browsers.

It allows scripting in several languages like Java, C#, PHP and Python.

Assertion statements in selenium provide an efficient way of comparing expected and actual results.

Selenium consist of following components

1. Selenium IDE

2. Selenium RC

3. Selenium Grid

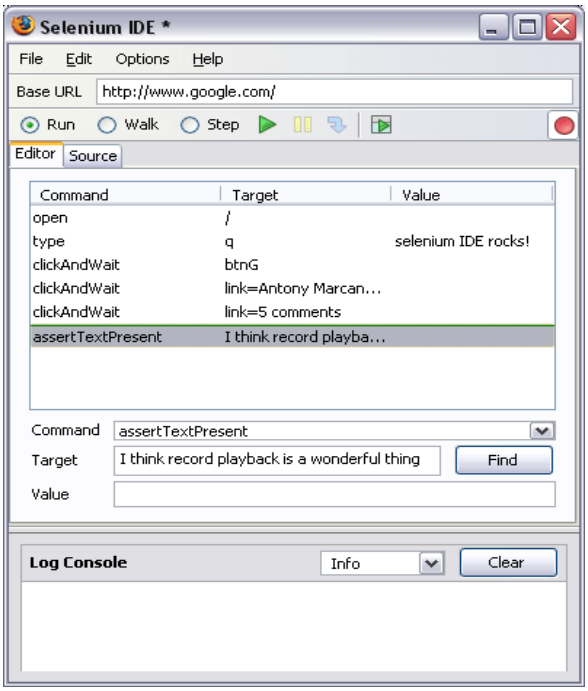


Figure1 Selenium IDE-UI

3.2 *Selenium IDE*

- Selenium IDE is an integrated development environment for Selenium tests. It is implemented as a Firefox extension, and allows you to record, edit, and playback the test. Selenium IDE
- Provides a way to save tests as Java, Ruby scripts, HTML or any other format.
- To use selenium tools first we have to install it
- Using Firefox, first download the IDE from the Selenium HQ downloads page then Restart Firefox. After that you will find the Selenium-IDE listed under the Firefox Tools menu.

Selenium allows you to

- To set the Text Encoding Format
- Setting of Default Page Time Out
- Provides the Base URL recording option
- Adding Selenium Core and IDE Extension

3.3 *Procedure of creating test cases [5]*

- For recording and playing test cases first
- Open Firefox which has IDE installed
- Open the application to record.
- Go To Tools → and opened the Selenium IDE
- Now perform the various operations on the application as you are testing the application.
- Once your recording is done click on the stop recording button and save the test case through the file menu. By default it will be saved in HTML format.
- To play the recorded test Click on the run button which is present on the user interface and you can also manage the speed of the execution.
- Once test execution is over you can view the test result in the bottom of the IDE window

3.4 *Selenium RC (Remote Control)*

- Selenium RC is a solution to cross browser testing.
- It is a server which is written in Java and available on all the platforms.

3.5 *Selenium Grid*

- Selenium-Grid allows the test suites to be run in multiple environments.
- Multiple instances of Selenium-RC are running on various operating system and browser configurations with the help of selenium-grid.

4.Conclusion

- Test automation process is use to minimize the cost and other overheads. Automation results in reduction of time spend on regression tests and provide an opportunity for organizations to improve the quality of their software products. Because automation tool can perform test faster than human.
- But effectiveness of automation testing depends on selection of appropriate tool which is compatible with checklist, selection of test cases to automate. Cost is also an important factor considered for selection of tool.
- Selenium tests is use to check workflow but it cannot be put in development build and it cannot be considered as replacement for exploratory testing

2. ACKNOWLEDGMENTS

We are very thankful to uday patkar who have contributed towards development of this template.. and other teaching staff who helps in the development of this paper

3. REFERENCES

[1] The Application of Software Testing Technology on Security in Web Application System" ISSN: 1662:7482 Vals-556-562 Department of information Engineering, Henan polytechnic China 450046.

[2] Research on Software Testing Tools Rasneet Kaur ChauhanȦ* and Iqbal SinghB Ȧ Department of Computer Science India

[3] Software Test Automation Abdul Rauf EMa,*, E.Madhusudhana Reddyb aResearch and Development Centre ,Bharathiyar University,Coimbatore-641014,

[4] Achievements for software testing Mark Harman, Yue Jia and Yuanyuan ZhangUniversity College London, CREST Centre, London, UK

[5] Software Testing Research and Practice ISTI-CNR Italy

International Journal of Computer Applications Technology and Research
Volume 5– Issue 4, 202 - 209, 2016, ISSN:- 2319–8656

Usage of Self-Organizing Map for Clustering Vertices

Rashmi Lad
MIT, Arts Commerce and Science College, Alandi (D), Pune-412106, M. S., India

P S Metkewar
Symbiosis Institute of Computer,Studies and Research (SICSR) Pune-411016, M. S., India

R.S. Walse
College of Dairy technology, Pusad, Nagpur, M.S., India

Abstract – Usage of Self-Organizing Map (SOM) for clustering vertices of any given graph. Simultaneously its input is observed and worked in terms of weight matrix, learning rate and final resultant matrix, which helps to form a cluster. The purpose of this paper is to introduce a procedure of SOM for clustering and observed impact corresponding to varied weight class for simple graph or vector matrix using Euclidean distance. A simple vector matrix problem is solved by using 2, 3 & 4 weight class matrix. By adopting a different weight matrix class with same vector matrix has presented a clustering and visualization.

Keywords – self-organizing map, topology, visualization, clustering, Euclidean distance.

INTRODUCTION

Self-Organizing Map (SOM) was developed by Professor Kohonen's. It is also called as Kohonen's SOM (Self Organizing map).SOM works on unsupervised learning which means training without any guidance or teacher. SOM learns on its own from beginning till the end and it is unsupervised competitive learning. The self-organizing Map is a special type of neural network that accepts N-dimensional input vector and maps them, in which neurons are organized in a hexagonal or rectangular grid and find the feature space with its neighboring neuron.

The main objective of this learning algorithm is that the network forms the feature map which takes input data and maps them into 1 or 2-dimensional feature space. The main feature of SOM is that it contains only two layers, the input layer, and an output layer. There is no hidden layer in SOM so it is different from other learning algorithms like feedforward back prorogation learning algorithm.

LITERATURE REVIEW

Authors [1] have observed that "On the use of Self-Organizing Map for clustering and visualization" of the number of output units. In this paper, SOMs can be used for clustering and visualization separately or simultaneously. There are various types of application used to compare SOM with other statistical approaches.

Authors [2] have focused "Clustering of the Self-Organizing Map" with different approaches. Authors used hierarchical agglomerative clustering and partitive clustering using K-means. By using SOM, they produced the prototypes and then performed direct clustering of the data and to reduce the computation time.

Authors [3] have observed that "Clustering Application of SOM neural network in clustering" is an unsupervised neural network for two-dimensional maps. It finds the similar data that will map to nearby locations. In this paper, authors introduce an experiment to analyze the SOM in clustering.

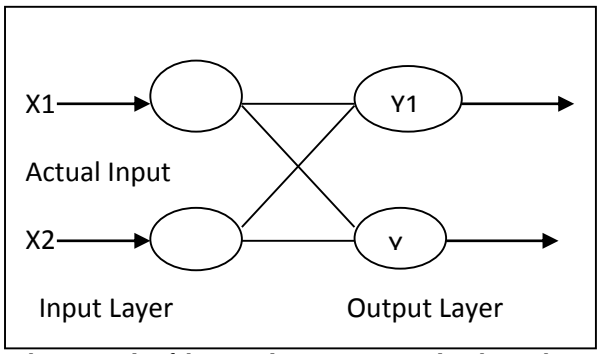
Authors [4] have emphasized that "The use of Three-dimensional Self-Organizing Maps for Visualizing Clusters in Geo-referenced Data" that maps 3D SOM data for visualizing clusters in geo-referenced data. This paper provides a comparison of a 2D or 3D SOM for a problem and increases the clustering quality of 3D SOMs.

In this paper Section 1 gives a review of the problem definition. Section 1.1 presents the architecture of self-organizing map. Section 1.2 describes the methodology of self-organizing map Section 1.3 shows Pseudo code section 1.4 describe how to train SOM and Section 1.5 shows research methodology. Section 2 it describes the computational analysis of self-organizing map. Section 2.1 describes the computational result analysis of self-organizing. Section 2.1.1 gives the result based on 2 weight class matrix. Section 2.1.2 describes the result based on 3 weight class matrix. Section 2.1.3 describes the result based on 4 weight class matrix. Section 3 – Describe the result analysis that is available after the large calculation.

1. PROBLEM DEFINITION

The Self-Organizing Map (SOM) is an artificial neural network that is very effective for clustering via visualization. It is very difficult to visualize SOM for the vector data. In this paper, we show that the vertices of SOM can be used successfully for visualizing clusters using Euclidean distance method of the neural network. We try to find the small distance between nearest cluster. In this paper, we use the vector matrix and 2, 3 and 4 weight matrix classes to find the winning neuron or the resultant output layer. Based on this we can find nearest cluster group and observed the changes.

1.1 Architecture of SOM



Each input node of the input layer is associated with weight w_{ij} that is adjusted during training. The SOM maintains

topological relationships between inputs in such a way that the neighboring inputs in the input layer are associated with neighboring neurons.

1.2. Pseudo Code

^ Declare n, m and Set its value by 5 and 1

^ Set learning rate by default 0.5

^ Declare input vector matrix

^ Declare weight class matrix

^ Repeat the procedure until m is less than or equal to n
Repeat the procedure for 1 to 5 for n
Check when m equal to 1
Store first input matrix
Check when m equal to 2
Store second input matrix
Check when m equal to 3
Store third input matrix
Check when m equal to 4
Store fourth input matrix
Otherwise
Store fifth input matrix
Stop
^ Repeat the procedure for 1 to 5

Sum ← Sum + square root (weight matrix value – input matrix value)

Sum = Square root of(sum);

^ Find min sum

^ Update that weight matrix (new) = weight matrix (old) + learning rate *(input matrix – weight matrix)

^ learning rate ← learning rate -0.1

^ Stop

1.3. Training SOM

There are two types of operation in self-organizing map

1. **Training Phase**

In the Training phase, the output node is found with the help of Euclidean distance between the input vector and the weight class connecting to that input and finds the minimum between them. This node is called the winner node and weight class. Now the weight of the neighboring output node will be updated so that the new weight is closer to the current input vector. This procedure is repeated for all input vectors and weight till they become constant. After one iteration or epoch of input vector, the learning rate gets changed and is multiplied by 0.5 at every epoch. In this way after applying the input vector, only the winner unit is determined.

This function is selected for the size of weight change in the distance of the neuron. This distance is calculated with the topology defined on the output layer of the network.

2. **Clustering Phase**
After training the SOM should give visualization where similar data are clustered within close proximity, and having smooth transitions or overlaps where clusters change.

1.4. Research Methodology

Exploratory research is one type of research method which is based on the theoretical idea. Researcher gets an idea from currently available theory and tries to elaborate or understand more about that topic. Sometimes it is the initial groundwork for this type of research. Exploratory research used in two ways; either a new topic or a new idea. A new topic is finding from the currently existing theory. New idea can come to understand exiting theory and set new perception according to that.

2. Computational Analysis of SOM

The Experiment is derived by using following directed graph

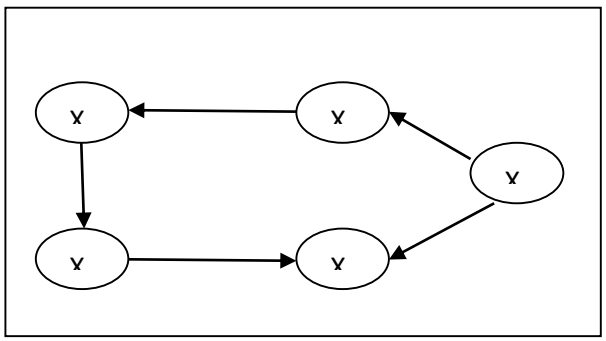


Figure 1: Directed graph

There is 5 node or input in the directed graph. They are connected with each other. The representations of Adjacency matrix (5 X 5) are as follows.

	X1	X2	X3	X4	X5
X1	0	1	0	0	0
X2	0	0	0	0	1
X3	0	0	0	1	1
X4	1	0	0	0	0
X5	0	0	0	0	0

Weight matrix for input is 4X5 which means there are 4 classes for 5 input nodes. The weight initializes between 0 to 1 only.

To evaluate the performance of the proposed algorithm, we tested 5 input vector and 3 different classes. The SOM algorithm was tested 5 input vector and 3 different classes. The SOM algorithm was executed using DOTNET code. The source and input parameters for these problems are shown.

Initial learning rate $\eta = 0.5$

After initialize learning rate and weight matrix calculate Euclidean distance with the following formula.

$$D(j) = \text{sqrt}\left(\sum_{i=1}^{n} (X_i - w_{ij})^2\right)$$

Where n is no of the input node. To find the distance between input vector subtract the input vector with weight matrix value and find the square root of that. After this calculates the summation of all input values for D1.value of j varies according to the number of vectors of weight matrix.

In this way repeat the same procedure for all the classes of weight matrix. Then compare all distances and find the minimum distance between them. That minimum distance is the Best Matching Unit (BMU) or winning neuron.

The distance of that weight matrix class is BMU or winning neuron. Now update the weight matrix on the winning cluster with the following formula.

$$W_{ij}(new) = W_{ij}(old) + \eta * [x_i - w_{ij}(old)]$$

Then get the new or updated weight matrix. Follow the same procedure for all input vectors. After repeating the same procedure for all input vectors, **one epoch** is over. The learning rate will change after one epoch. Learning rate will decrease with the following formula.

$$\eta = \eta * 0.5$$

The default learning rate is 0.5.Now repeat the same for another epoch till updated weight matrix does not get a similar result.

2.1. Computational Result Analysis of SOM

A summary of all SOM parameters used for solving the problems is given in the table.

2.1.1 Input vector 5X5 and weight class 2

Class means a set or category of things having some property or attribute in common and differentiated from others by kind, type, or quality.

Weight Matrix for class 2

Wij =

	w_{11}	w_{12}	w_{13}	w_{14}	w_{15}
w1j	.2	.4	.6	.8	1
w2j	.9	.7	.5	.3	.1
	w_{21}	w_{22}	w_{23}	w_{24}	w_{25}

Epoch 1- when η=0.5

In this example first take input vector v1 & weight of w1j (where j = 1 to 5) and find the distance1. Perform similar procedure for same input vector v1 & another weight class w2j (where j =1 to 5) and find distance 2.Now check minimum between Dist1 and Dist2. When we get minimum distance then update weight matrix as per the winning neuron. If dist1 is minimum then w1j will change otherwise w2j class. Now the similar procedure is followed for the entire input vector v2, v3, v4 and v5 and the updated weight matrix is found.

Table 1: Result of weight matrix class 2 with learning rate 0.5

	Input vector	Winning Neuron	Dist 1	Dist 2	Mini mum	Updated weight	
						Δw1	Δw2
Iteration 1	v1 =[0,1,0,0,0]	B	2.4	1.25	1.25	No change	[.45,.85,.25,.15,.05]
Iteration 2	v2 =[0,0,0,0,1]	A	1.2	1.91	1.2	[.1,.2,.3,.4,1]	No change
Iteration 3	v3=[0,0,0,1,1]	A	0.5	2.61	0.5	[.05,.1,.15,.7,1]	No change
Iteration 4	v4=[1,0,0,0,0]	B	2.42	1.11	1.11	No change	[.725,.425,.125,.075,.025]
Iteration 5	v5=[0,0,0,0,0]	B	1.52	0.72	0.72	No change	[0.36,.213,.062,.038,.013]

When one epoch is over so decrease the learning rate with 0.1.Then again same procedure is followed for the input vector v1 to v5.

Table 2: Result of weight matrix class 2 with learning rate 0.4(Epoch 2)

	Input vector	Winning Neuron	Dist1	Dist2	Minimum	Updated weight	
						Δw1	Δw2
Iteration 1	v1 =[0,1,0,0,0]	B	2.32	0.75	0.75	No Change	[.218,.528,.038,.023,.007]
Iteration 2	v2 =[0,0,0,0,1]	A	0.52	1.31	.052	[.03,.06,.09,.42,1]	No change
Iteration 3	v3=[0,0,0,1,1]	A	0.34	2.26	0.34	[.018,.036,.054,.65,1]	No change
Iteration 4	v4=[1,0,0,0,0]	B	2.39	0.89	0.89	No change	[.531,.317,.023,.014,.004]
Iteration 5	v5=[0,0,0,0,0]	B	1.52	0.72	0.72	No change	[0.318,.19,.014,.014,.003]

After repeating the same procedure with decreasing learning rates, the result will be found zero (η=0) after 6 epoch and the values of all the input vectors will be same in class1 and class 2**.**

Table 3: Result of weight matrix class 2 with learning rate 0.0(Epoch 6)

	Input vector	Winning Neuron	Dist1	Dist2	Minimum	Updated weight	
						Δw1	Δw2
Iteration 1	v1 =[0,1,0,0,0]	B	2.32	0.66	0.66	No Change	[.326,.253,.002,.001,0]
Iteration 2	v2 =[0,0,0,0,1]	A	0.34	1.61	0.34	[.005,.009,.014,.583,1]	No change
Iteration 3	v3=[0,0,0,1,1]	A	0.17	2.16	0.17	[.005,.009,.014,.583,1]	No change
Iteration 4	v4=[1,0,0,0,0]	B	2.33	0.51	0.51	No change	[.326,.253,.002,.001,0]
Iteration 5	v5=[0,0,0,0,0]	B	1.34	0.17	0.17	No change	[.326,.253,.002,.001,0]

Map Decision for 5X5 input matrix and 2 classes

Here A and B based on weight class. In this example there are 2 weight classes so that in each epoch for v1 B is the winning neuron for v2 A is the winning neuron, for v3 A, for v4 B, and for v5 B is the winning neuron.

Table 4: Map table of input matrix and weight class 2

	A	B
v1	0	1
v2	1	0
v3	1	0
v4	0	1
v5	0	1

Here the input matrix is 5X5 and weight matrix is 5X2, it will be reducing in 5X2 output matrix and there will be three clusters only. c1 = {v2, v3}, c2 = {v4, v5} and c3 = {v1} have similar distance or nearest distance with each other.

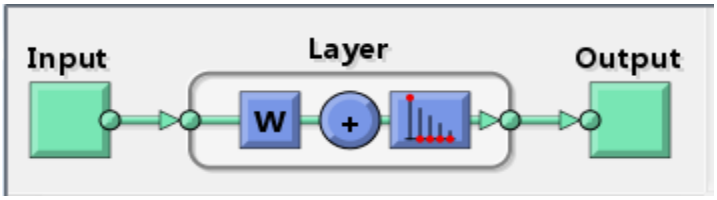


Figure 2: Architecture diagram of map table for weight class 2

2.1.2 Input vector 5X5 and weight class 3

Weight Matrix for class 3

Wij =

	J = 1 to 5				
w1j	.2	.4	.6	.8	1
w2j	.9	.7	.5	.3	.1
w3j	.3	.6	.9	.2	.4

In this example first take input vector v1 and weight of w1j (where j = 1 to 5) and find the distance1. Similar procedure performed for same input vector v1 with another weight class w2j (where j =1 to 5) and with third weight class wj3. After that find minimum distance between dist1, dist2 and dist3 as per the formula and find the new weight that is called updated weight for the second input vector. If dist1 is minimum then updated weight class will change Δw1, if dist2 minimum then updated weight class Δw2 will changed otherwise Δw3 will be changed.

Now the similar procedure is followed for the entire input vector v2, v3, v4 and v5 and finds the updated weight matrix.

Table 5: Result of weight matrix class 3 with learning rate 0.5(Epoch 1)

	Input vector	**Winning Neuron**	**dist1**	**dist2**	**dist3**	**Min**	**Updated Weight**		
							Δw1	Δw1	Δw3
Iteration 1	v1=[0,1,0,0,0]	B	2.4	1.25	1.26	1.25	No Change	[.45,.85,.25,.15,.05]	No Change
Iteration 2	v2=[0,0,0,0,1]	A	1.2	1.912	1.66	1.2	[.1,.2,.3,.4,1]	No Change	No Change
Iteration 3	v3=[0,0,0,1,1]	A	0.5	2.6	2.26	0.5	[.05,.1,.15,.7,1]	No Change	No Change
Iteration 4	v4=[1,0,0,0,0]	B	2.4	1.112	1.8	1.112	No Change	[.725,.425,.125,.075,.025]	No Change
Iteration 5	v5=[0,0,0,0,0]	B	1.5	0.72	1.4	0.72	No Change	[.362,.212,.062,.037,.012]	No Change

After completing one epoch same procedure will be followed for the next epoch and learning rate will be decreased by 0.1, this procedure will continue till learning rate becomes zero. Because at this stage weighted weight get a constant value.

Table 6: Result of weight matrix class 3 with learning rate 0.4(Epoch 2)

	Input vector	**Winning Neuron**	**out1**	**out2**	**out3**	**Min**	**Updated Weight**		
							Δw1	Δw1	Δw3
Iteration 1	v1 =[0,1,0,0,0]	B	2.3	0.75	1.26	0.75	No Change	[.217,.527,.037,.022,.007]	No Change
Iteration 2	v2 =[0,0,0,0,1]	A	0.5	1.31	1.66	0.5	[.03,.06,.09,.42,1]	No Change	No Change
Iteration 3	v3=[0,0,0,1,1]	A	0.34	2.26	2.26	0.34	[.018,.036,.054,.652,1]	No Change	No Change
Iteration 4	v4=[1,0,0,0,0]	B	2.3	.89	1.8	.89	No Change	[.530,.316,.022,.013,.004]	No Change
Iteration 5	v5=[0,0,0,0,0]	B	1.4	0.38	1.4	0.38	No Change	[.318,.189,.013,.008,.002]	No Change

Table 7: Result of weight matrix class 3 with learning rate 0.0(Epoch 6)

	Input vector	**Winning Neuron**	**out1**	**out2**	**out3**	**Min**	**Updated Weight**		
							Δw1	Δw1	Δw3
Iteration 1	v1 =[0,1,0,0,0]	B	2.3	.66	1.26	.66	No Change	[.325,.253,.001,.001,.000]	No Change
Iteration 2	v2 =[0,0,0,0,1]	A	.34	1.16	1.66	.34	[.004,.009,.013,.583,1]	No Change	No Change
Iteration 3	v3=[0,0,0,1,1]	A	.17	2.16	2.26	.17	No Change	No Change	No Change
Iteration 4	v4=[1,0,0,0,0]	B	2.33	.51	1.8	.51	No Change	No Change	No Change
Iteration 5	v5=[0,0,0,0,0]	B	1.34	.17	1.4	.17	No Change	No Change	No Change

After repeating the same procedure with decreasing learning rates, the result will be found the values of all the input vectors will be same in class1, class 2and class 3**.**

Map Decision for 5X5 input matrix and 3 classes

Here A and B based on weight class. In this example there are 3 weight classes so that in each epoch for v1 B is the winning neuron for v2 A is the winning neuron, for v3 A, for v4 B and for v5 B is the winning neuron. The third weight class is constant.

Table 8: Map table of input matrix and weight class 3

	A	B
v1	0	1
v2	1	0
v3	1	0
v4	0	1
v5	0	1

Here the input matrix is 5X5 and weight matrix is 5X3, it will be reducing in 5X2 output matrix and there will be one cluster only. c1 = {v2, v3}, c2 = {v4, v5} c3 = {v1} have similar distance or nearest distance with each other.

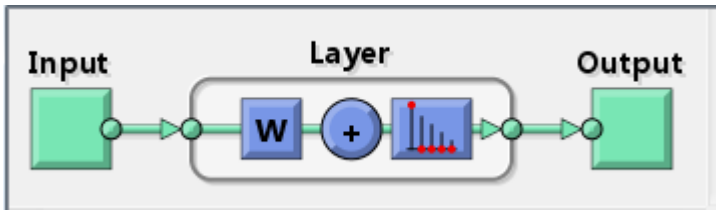


Figure 3: Architecture diagram of map table for weight class 3

2.1.3 For input vector 5X5 and weight class 4

Weight Matrix for class 4

Wij =

	j = 1 to 5				
w1j	.2	.3	.7	.8	.9
w2j	.9	.8	.6	.3	.1
w3j	.3	.6	.7	.2	.4
w4j	.1	.3	.4	.6	.9

In this example first take input vector v1 and weight of w1j (where j = 1 to 5) and find the distance1. Similar procedure performed for same input vector v1 with another weight class w2j (where j =1 to 5), with third weight class wj3 and w4j also. After that find minimum distance between dist1, dist2, dist3 and dist4 as per the formula and find the new weight that is called updated weight for the second input vector. If dist1 is minimum then updated weight class will change Δw1, if dist2 minimum then updated weight class Δw2 will change, if dist3 minimum then updated weight class Δw3 will change otherwise Δw4 will change.

Now the similar procedure is followed for the entire input vector v2, v3, v4 and v5 and finds the updated weight matrix.

Table 9: Result of weight matrix class 4 with learning rate 0.5(Epoch 1)

	Input vector	**Winning Neuron**	**out1**	**out2**	**out3**	**out4**	**Min dist**	**Updated Weight**			
								Δw1	Δw1	Δw3	Δw4
Iteration 1	v1 =[0,1,0,0,0]	C	2.4	1.3	.94	1.7	1.3	No Change	No Change	[.15,.8,.35,.1 ,.2]	No Change
Iteration 2	v2 =[0,0,0,0,1]	A	1.2	2.7	1.43	1.44	1.2	[.1,.15,.35,.4,.9 5]	No Change	No Change	No Change
Iteration 3	v3=[0,0,0,1,1]	A	.5	3.1	2.2	2.24	.5	[.05,.08,.17,.7,. 98]	No Change	No Change	No Change
Iteration 4	v4=[1,0,0,0,0]	B	2.3	1.11	1.5	2.34	1.11	No Change	[.95,.4,.3,.15,.0 5]	No Change	No Change
Iteration 5	v5=[0,0,0,0,0]	C	1.4	1.17	.83	1.64	.83	No Change	No Change	[.08,.4,.17,.0 5,.1]	No Change

After completing one epoch same procedure will be followed for the next epoch and learning rate will be decreased by 0.1, this procedure will continue till learning rate becomes zero. Because at this stage weighted weight get the constant value.

Table 10: Result of weight matrix class 4 with learning rate 0.4(Epoch 2)

	Input vector	Winning Neuron	out 1	out 2	out 3	out 4	Min dist	Updated Weight			
								Δw1	Δw1	Δw3	Δw4
Iteration 1	v1 =[0,1,0,0,0]	C	2.3	1.3	.40	1.2	.40	No change	No Change	[.05,.64,.1,.03,.06]	No Change
Iteration 2	v2 =[0,0,0,0,1]	A	.52	2.0	1.3	1.31	.52	[.03,.05,.1,.4 2,.99]	No Change	No Change	No Change
Iteration 3	v3=[0,0,0,1,1]	A	.35	2.7	2.2	2.25	.35	[.02,.03,.06,. 65,.99]	No Change	No Change	No Change
Iteration 4	v4=[1,0,0,0,0]	B	2.3	.27	1.3	2.14	.277	No Change	[.97,.24,.18 ,.09,.03]	No Change	No Change
Iteration 5	v5=[0,0,0,0,0]	C	1.4	1.0	.42	1.23	.42	No Change	No Change	[.03,.38,.06,.02,.0 4]	No Change

Table 11: Result of weight matrix class 4 with learning rate 0.0(Epoch 6)

	Input vector	Winning Neuron	out 1	out 2	out 3	Out4	Min	Updated Weight			
								Δw1	Δw2	Δw3	Δw4
Iteration 1	v1 = [0,1,0,0,0]	C	2.33	1.72	.34	1.15	.34	No Change	No Change	[.01,.43,.02,.0 ,.01]	No Change
Iteration 2	v2 = [0,0,0,0,1]	A	.35	1.9	1.2	1.21	.35	[.01,.01,.02,. 58,1]	No Change	No Change	No Change
Iteration 3	v3= [0,0,0,1,1]	A	.21	2.8	2.1	2.20	.21	No Change	No Change	No Change	No Change
Iteration 4	v4= [1,0,0,0,0]	B	2.3	.03	1.2	2.01	.03	No Change	[.98,.12,.09,.05 ,.02]	No Change	No Change
Iteration 5	v5=[0,0,0,0 ,0]	C	1.3	.99	.22	1.03	.22	No Change	No Change	[.01,.43,.02,.0 ,.01]	No Change

After repeating the same procedure with decreasing the learning rate, the result will be found and the values of all the input vectors will be same in class1, class 2, class 3 and class 4.

Map Decision for 5X5 input matrix and 4 classes

Here A, B, C and D based on weight class. In this example there are 4 weight classes so that in each epoch for v1 C is the winning neuron for v2 A is the winning neuron, for v3 A, for v4 B and for v5 C is the winning neuron.

Table 12: Map table of input matrix and weight class 4

	A	B	C
v1	0	0	1
v2	1	0	0
v3	1	0	0
v4	0	1	0
v5	0	0	1

In every epoch, the BMU or winning neuron will be same. As per this discussion the map is like this.
Here the 5X5 will be reducing in 5X3 output matrix and there are two clusters are.
c1 = {v1}, c2 = {v2, v3}, c3={v4} and c4 ={v5} have similar distance or nearest distance with each other.

Architecture of Output matrix for 4 classes SOM

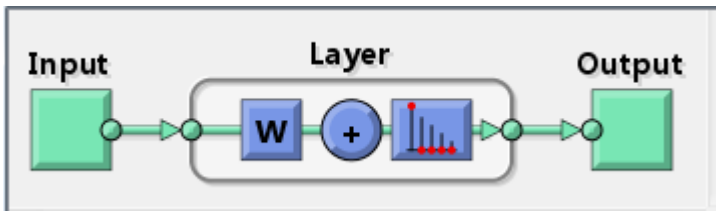


Figure 4: Architecture diagram of map table for weight class 4

3. RESULT ANALYSIS

The main aim of clustering is to reduce the data by grouping or categorizing them. There are different types of clustering methods. But here we apply partition clustering method. Clustering is used to reduce the data and to make a categorization. Partitioning cluster directly decomposes the data into a disjoint cluster.

In these problems, there are 5X5 input vectors. But every time on applying different weight classes on the same input vector, if the classes are n (n > 2) the result will come for only n-1 classes and the result of last class will be constant in each iteration and in each epoch.

Table 13: Topological mapping of all weight classes

Input Vector	Weight classes	Topology Map(output layer)
5x5	**5x2**	**5x2**
5x5	**5x3**	**5x2**
5x5	**5x4**	**5x3**

The topological neighboring decline monotonically, from a value less than half the largest diagonal of the map. **This is necessary condition for convergence**

Result analysis of this problem is that the topology size is not equivalent to the size of weight matrix. If weight size will be the increase there will be the change in topology.

CONCLUSION

In this paper, we observed that learning rates the change after every epoch. If learning rate is constant then the result could not be found or we cannot visualize the cluster. When learning rate becomes zero then only the value of weight matrix gets constant. The value of learning rate changes either by decreasing it by 0.1 or multiplying it by 0.5. In every epoch the winning neuron is same.

Moreover, in this paper, we also observed that when we calculate the distance for all vectors using one learning rate than one epoch is over. But for one epoch many iterations is performed. When we calculate the distance between one vector it is called iteration and when the same iteration is repeated for all vector it is called epoch.

Here we used different types of weight matrix such as 2, 3 and 4 for the same one-dimensional array. The value of weight matrix is changed based on iteration or the distance, we calculate the new weight for that class which has the minimum distance.

Thus, one can conclude that in the case of 2 class weight matrixes, we get only 2 output layers or winning neuron. For 3 class weight matrix we get only 2 output layers or winning neuron and for 4 class weight matrix, we get only 3 output layers or winning neuron. Here we observed that if weight matrix size is increased than one class is constant and we get output layer always minus one from weight matrix.

FUTURE WORK

Future work may further investigate on large data which has meaningful data items in the sets and find the variation on large data sets also. We will try to take multi-dimensional meaning full data and find the nearest clusters and minimize the data in terms of rows & columns.

In future work we investigate the effect of increase the weight matrix class on multi-dimensional data also. We will also work on learning rate and try to find the cluster to reduces the epoch and iterations and minimize the calculation.

REFERENCES:

1] “On the use of self-organizing maps for clustering and visualization” by Arthur Flexer.

[2] “Clustering of the Self-Organizing Map” by Juha Vesanto and Esa Alhoniemi IEEE transactions on neural networks, vol. 11, no. 3, pp 586-600, may 2000.

[3]”Application of SOM neural network in clustering” by Soroor Behbahani, Ali Moti Nasrabadi J. Biomedical Science and Engineering, 2009, 2, 637-643.

[4] “On the use of Three-dimensional Self-Organizing Maps for Visualizing Clusters in Geo-referenced Data” by Jorge M. L. Gorricha and Victor J. A. S. Lobo

[5] Application of Visual Clustering Properties of Self-Organizing Map in Machine-part Cell Formation Manojit Chattopadhyay, Pranab K. Dan, Sitanath Majumdar.

[6] “Improving Performance of Self-Organising Maps with Distance Metric Learning Method” by Piotr P lo´nski and Krzysztof Zaremba published 1407.1201v1 [cs.LG] 4 Jul 2014.

[7] “Clustering Internet Usage Behaviours with SOM Neural Networks” by U. Celenk, O. Ucan Proceedings of the World Congress on Engineering and Computer Science 2012 Vol II WCECS 2012, October 24-26, 2012.

[8] “Self-Organizing Map -based Document Clustering UsingWordNet Ontologies” by Tarek F. Gharib, Mohammed M. Fouad, Abdulfattah Mashat, Ibrahim Bidawi IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012 ISSN (Online): 1694-0814.

[9] “Clustering with SOM: u*c” by Alfred ultsch.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 4, 210 - 214, 2016, ISSN:- 2319–8656

Powerful Combination of Color Descriptor and LBP Descriptor for Image Retrieval

Nawal Chifa
EEA&TI laboratory, Hassan II University of Casablanca
Faculty of Sciences and Techniques (FSTM)
Mohammedia, Morocco

Abdelmajid Badri
EEA&TI laboratory, Hassan II University of Casablanca
Faculty of Sciences and Techniques (FSTM)
Mohammedia, Morocco

Yassine Ruichek
Syst&Transp Lab, Univ. of Technoogyl. UTBM-90010
Belfort Cedex, France

Aicha Sahel
EEA&TI laboratory, Hassan II University of Casablanca
Faculty of Sciences and Techniques (FSTM)
Mohammedia, Morocco

Khadija Safi
EEA&TI laboratory, Hassan II University of Casablanca
Faculty of Sciences and Techniques (FSTM)
Mohammedia, Morocco

Abstract: The search for visual information in large mass of multimedia data has become essential with the digital evolution. This sparked a need for development of information search techniques by visual content; the performance of such a search system depends largely on the choice of descriptors and technical employees of their extractions. In our work, we present techniques for extracting local and global descriptors applied to two bases of different images, with a connection between the global and local descriptors approach, performed on the two bases of images, followed by a comparative study of different methods used.

Keywords: CBIR, combinations of descriptors, global and local descriptor, Histogram HSV color, LBP descriptor.

1. INTRODUCTION

The development of an image by visual content search system to be effective for large collections of images requires expertise in both image analysis and database management. Such a system is used to characterize images by visual descriptors and search for these images by similarity from these descriptors. Figure1 illustrates this mechanism.

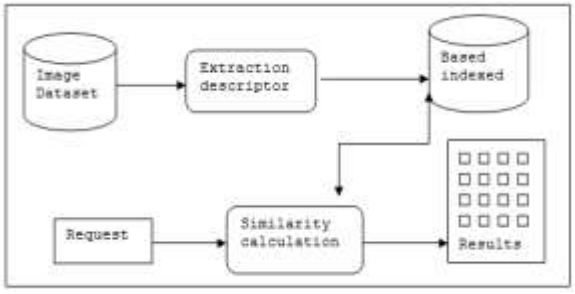


Fig1: Architecture of an image search system.

Image search systems usually deal with extracting visual features, often the color, shape and texture; we use these features in order to provide a comprehensive description of the images [1]. However, these fail when one we want to consider the semantics of the objects described in the image [1] [2].

To overcome this problem, in the following work, we will focus on the extraction of global descriptors images, precisely the color HSV and a local descriptor (operator of local binary patterns) while exploring different methods of connections between them and by conducting a comparative study of these methods on the results.

2. TECHNIQUES AND METHODS USED

Several extraction methods of visual descriptors have been proposed for visual recognition. These image description methods often depend on the applications used. Some distinguishes descriptors reflecting the overall visual appearance of an image, such as color histogram [3], color moments [4], the co-occurrence matrix [5], edge histogram [6], and so on. These features are extracted from the whole of an image and don't give information for the specific region of image. In These global features, unlike in the comprehensive local approaches, methods of local description are intended to describe the content of the image locally. They thus offer the ability to perform a search to on a part of the image or on an object present in the image. The idea of local image descriptors is to extract features from local image region center , This approach involves cutting or segmenting the image into regions of interest , or to determine the points of interest, such as SIFT[7], PCA-SIFT[8],SURF [9],the local binary pattern (LBP) operator[10].

2.1 The data bases used:

In our study, we used two image databases; the first one is a gathering of nature scenes classified according to several themes: The Simplicity dataset is a subset of COREL image dataset. It contains a total of 1000 images, which are equally divided into 10 different categories (Figure2), and the second database contains 810 texture images from nine materials KTH-TIPS-b dataset (Figure 3):

Fig2: The Simplicity dataset is a subset of COREL

Fig3: KTH-TIPS-b dataset

2.2 HSV color histogram:

In our system, the database images are color images. Algorithms calculate histograms colors are easy to implement with a very short turnaround time, introducing invariance to rotation and translation. However, these histograms have no spatial information on the colors of the positions [11]. To overcome this problem we used an image division method for extracting aggregate information partially and collect them later in the same order e(left to right ant top to down) figure4.

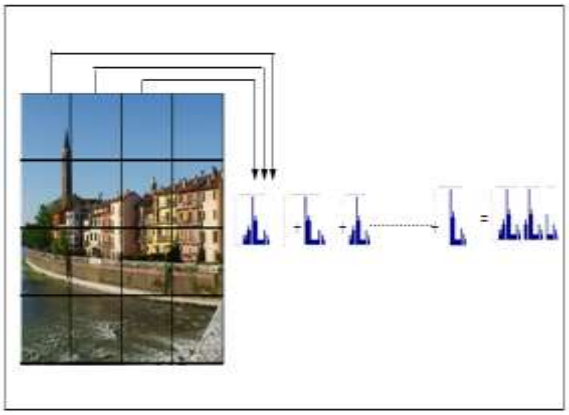


Fig4: Division image to block and extract HSV histogram from each block, and concatenate all of them.

Since the color histogram is sensitive to small changes in brightness, which is problematic if we want to compare similar images, acquired under different conditions, we opted for the HSV color space and we have merged (merged what!!!!) with a descriptor local texture [12], as will be described in the following paragraphs.

2.3 Histogram local binary patterns:

The operator of the local binary patterns (LBP) was proposed in the late 90s by Ojala [13]. Extraction of LBP features is efficient and with the use of multi-scale filters; invariance to scaling and rotation can be achieve. The idea of this texture operator is to assign to each pixel a dependent code grayscale. The gray level of the center pixel (i_c) of coordinates (x_c, y_c) is compared with its neighbors (i_n) using the following equation (1). Figure 5 give an example:

$$\text{LBP}(x_c, y_c) = \sum_{n=0}^{p} s(i_n - i_c)$$
$$s(i_n - i_c) = 1 \text{ si } i_n - i_c \geq 0$$
$$= 0 \text{ si } i_n - i_c < 0 \qquad (1)$$

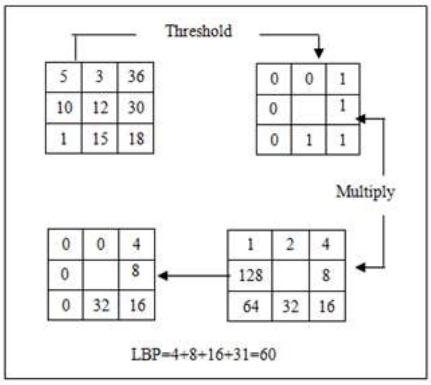


Fig.5

Example for calculation the LBP operator

Where p is the number of neighboring pixels. In general, we consider a neighborhood of 3 * 3 where p = 8 neighbors. So we get, as an image to grayscale, a matrix containing LBP values between 0 and 255 for each pixel. A histogram is calculated based on these values to form the LBP descriptor.

For our descriptor, we used the uniform LBP, which extracts the most fundamental structure from the LBP. A LBP descriptor is considered to be uniform if it has **at most** two *0-1* or *1-0* transitions. For example, the pattern 00001000 (2 transitions) and 10000000 (1 transition) are both considered to be **uniform patterns** since they contain at most two *0-1* and *1-0* transitions. The pattern 01010010 on the other hand is *not* considered a uniform pattern since it has six *0-1* or *1-0* transitions.

Based on this, we propose using those nine uniform patterns that have a U value of at most 2 (00000000, 00000001, 00000011, 00000111, 00001111, 00011111, 00111111, 01111111, and 11111111). These nine patterns correspond to 58 of the 256 original unrotated patterns that can occur in the

3x3 neighborhood. Remaining patterns are accumulated into a single bin, resulting in a 59-bin histogram.

Using only 58/256 of the pattern information may appear as a waste of information, but this approximation is supported by a very important observation. Namely, the chosen nine uniform patterns seem to contribute most of the spatial patterns present in deterministic micro-textures.

In order to obtain the color information with the uniform color LBP features, we calculate the uniform LBP descriptor independently over all the channels (H, S, V), and then concatenating them to get the color LBP, as like in figure 6.

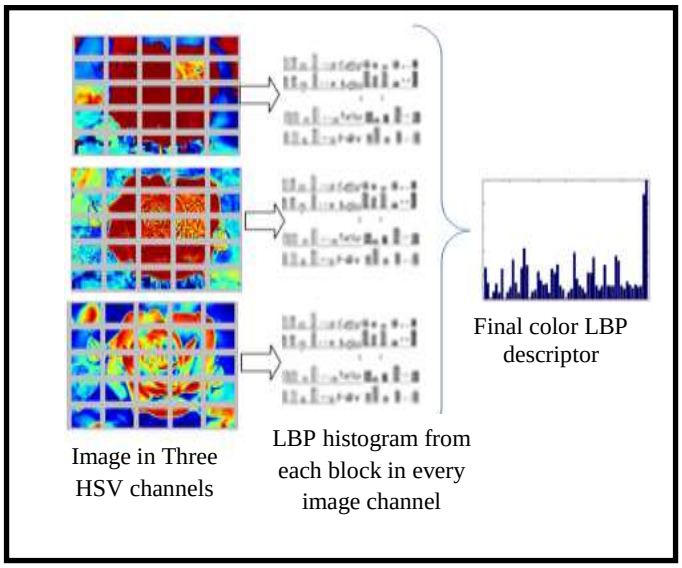


Fig6. Construction of local image descriptor with uniform LBP in three channels HSV

2.4 Combination of descriptors

First we tested and evaluated the results of each method, and this on both image databases, and then to overcome the limitations imposed by each descriptor, we combined the two methods by concatenating the vectors of descriptors and standardizing. And for the calculation of similarity between vectors we opted for the Euclidean distance which proved very optimal for comparing vectors and histograms [14].

- **Step1** : we divide the image to 16blocks for extract descriptor from each block:

- **Step2** : from each block we extract the histogram color and convert the same block to gray for extracting the histogram LBP and concatenate the two vectors

$$V_1 = Vlbp_1 + Vhsv_1$$

- **Step3**: loop over each block, left to right and the top to do down, and concatenate all the vectors in the same order to obtain the vector descriptor :

$$V_2 = Vlbp_2 + Vhsv_2$$
$$V_3 = Vlbp_3 + Vhsv_3$$
.
.
.
$$V_{16} = Vlbp_{16} + Vhsv_{16}$$

Combined descriptor= {V1, V_2, V_3 ...V_{16}}

3. Experimental results:

To evaluate our methods described above, we have set up an image search system that extracts the visual signatures of each image of the database as a vector of digital values and stores it in a data file. The signature of the query image will be compared later to those stored in the file according to the Euclidean distance, and return images with zero minimum distance to see the query image. To measure the quality of image search system content, parameters precision and recall are conventionally used [15]. Let Ai represents all relevant image results for a given query and Bi represents all the images result returned by the system. We define:

The precision as the ratio between the number of relevant images retrieved and number of images found:

$$Pi = (Ai \cap Bi)/ (Bi)$$

The recall as the number of relevant images found on the number of images relevant:

$$Ri = (Ai \cap Bi)/ (Ai)$$

Our system is designed to return 25 picture following a query image; for each query we calculate the average retrieval precision (ARP):

$$ARP = \frac{1}{N}\sum_{i=1}^{N} Pi$$

Where N is the size of testing category in dataset.

Table1 Comparison of the ARP values obtained by the proposed method with the standard Corel dataset Image

Descriptors	Block based colorLBP	Block based Histogram HSV	Block based HSV+LBP
Africa	0,7	0,72	**0,94**
Beach	0,3	0,5	**0,58**
Building	0,52	0,42	**0,61**
Bus	0,78	0,48	**0,96**
Dinosaur	1	0,98	**1**
Elephant	0,4	0,52	**0,61**
Flower	0,52	0,72	**0,78**
Horse	0,68	0,96	**0,98**
Mountain	0,54	0,38	**0,58**
Food	0,48	0,68	**0,78**
Average :	***59,2%***	***63,6%***	***78,2%***

Table 1 shows the precision rate for each method , we observe better performance for the combined descriptor on both databases (78%), however the individual use of LBP descriptor gives only (59,2%) and block HSV(63,6%).

Figure7 illustrates an example query image that found similar results for every method.

The same method was applied on the basis of the texture again. Our combination method has shown very effective results with an important average value (94%) compared to (60% and 77%) for the others methods extracting, as shown in table 2.

Table2: Comparison of the ARP values obtained by the proposed method with the texture dataset

Descriptors / Materials	Block based colorLBP	Block based Histogram HSV	Block based HSV+ULBP
Sandpaper	0,76	0 ,96	1
Aluminum	1	1	1
Styrofoam	0,42	0,9	0,98
Sponge	0,44	0,32	0 ,56
Corduroy	0,42	0,72	0 ,8
Linen	0,68	0,92	0,96
Brown bread	0,63	0,41	0,74
Cracker	0,45	0,9	0,92
Orange peel	0 ,72	1	1
AVERAGE	***60%***	***77%***	***94%***

An example for image retrieval using the three methods extraction is shown in figure7, we can see that the result using the LBP block does not respect the color distribution and the descriptor of color histogram does not respects shape of objects, against the combination of these two descriptors

gives satisfactory result in form and color

Fig7: Example of image result using the three methods

4. CONCLUSION AND PERSPECTIVES:

The histogram LBP and histogram color have no spatial information but in our algorithm we overcome this problem by using an image division method for extracting aggregate information partially and locally, furthermore a merger of descriptors carried out in our research system has shown a visible improvement rate in statistical data.

The effectiveness of a descriptor depends largely on the type of data and their heterogeneity, and the proposed combination in this work proved to be quite satisfactory and can give more performance on other types of base image.so it can be tested on other types of image-based to evaluate the performance of its results and bring him it further improvement by combining different kinds of descriptor and integrating indexing methods in our system.

5. REFERENCES

[1] M. Singha and K.Hemachandran Content Based Image Retrieval using Color and Texture Signal & Image Processing : An International Journal (SIPIJ) Vol.3, No.1, February 2012

[2] C. Liu and J. Yang « ICA color space for pattern recognition, » IEEE Trans. On Neural Networks, vol.20, no. 2, pp 248-257, 2009.

[3] M.J Swain, D.H Ballard, Color indexing, International Journal of Computer Vision 7 (1) (1991)11-32

[4] M.A Stricker, M.Orengo, Similarity of color image, in Proc. Of storage an Retrieval for Image and Video Databases, 1995, pp .381-392

[5]M. Tuceryan, A.K. Jain, Texture analysis, Handbook of pattern Recognition and Computer Vision, 2nd edition, World Scientific Publishing Co.,1998, pp.20-248

[6] D.K. Park, Y.S. Jeon, C.S. Won, Efficient use of local edge histogram descriptor, in Proc of ACM workshops on Multimedia, 2000, pp. 51-54.

[7] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91-110.

[8] Y. Ke, R. Sukthankar, PCA-SIFT: a more distinctive representation for local image descriptors, in Proc. of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2004, pp. 506-513.

[9]H. Bay, A. Ess, T. Tuytelaars, L.V. Gool, SURF: speeded up robust features, Computer Vision and Image Understanding 110 (3) (2008) 346-359.

[10] T. Ojala, M. Pietikainen, D. Harwood, A comparative study of texture measures with classification based on feature distribution, Pattern Recognition 29 (1996) 51-59.

[11] R. Brunelli, O. Mich : On the Use of Histograms for Image Retrieval, IEEE International Conference on

Multimedia Computing and Systems, vol. 2, p. 143-147(1999)

[12] OJALA T., PIETIKÂINEN M., MÂENPÂÂ T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Recognition. Vol. 24, Num. 7 (2002), 971–987

[13] W. Ben Soltana, A. Porebski, N. Vandenbroucke, A. Ahmad & D. Hamad: Contribution des descripteurs de texture LBP à la classification d'images de dentelles, Article RI 2014

[14] Herveé Jégou, Matthijs Douze, Cordelia Schmid: Exploiting descriptor distances for precise image search, [Research Report] RR-7656, INRIA. 2011-18pages

[15] Jing Yu a, ZengchangQin a,n, TaoWan b, XiZhang, « Feature integration analysis of bag-of-features model for image retrieval » Neurocomputing120(2013)355–364

International Journal of Computer Applications Technology and Research
Volume 5– Issue 4, 215 - 217, 2016, ISSN:- 2319–8656

A Review on Classification and Prediction Based Data Mining to Predict Slow Learners in Senior Secondary Schools

Sohajbir Singh Ubha

Rayat Bahra University, Mohali

Punjab, India

Gaganpreet Kaur Bhalla

Rayat Bahra University, Mohali

Punjab, India

Abstract: **One of the biggest challenges that higher education faces today is predicting the paths of students. Institutions would like to know, for example, which students will enrol in which course, and which students will need assistance in particular subject. The new interesting subject that is being offered by institution to interact more students. How management get better information about student, their result, about the success of new offered courses, the answer is data mining .it helps to institution to take decision .data mining is the better tool to predict the result of the student.**
In this paper, I will discuss about data mining their different phases, advantages of data mining in the field of academics, some results and tools used for data mining and their applications.
Keywords: KDD (Knowledge Discovery in Database), J48 Algorithm, Naïve Bayesian Classifier, WEKA data mining tool.

1. INTRODUCTION

We Data Mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help universities or institutions to focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing institution to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer institution questions that traditionally were too time consuming to resolve [5]. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining is a powerful tool for academic intervention. Through data mining, a university could, for example, predict with 85 percent accuracy which students will or will not graduate. The university could use this information to concentrate academic assistance on those students most at risk.

In order to understand how and why data mining works, it's important to understand a few fundamental concepts. First, data mining relies on four essential methods: classification, categorization, estimation, clustering and visualization [1]. Classification identifies associations and clusters, and separates subjects under study. Categorization uses rule induction algorithms to handle categorical outcomes, such as "persist" or "dropout," and "transfer" or "stay." Estimation includes predictive functions or likelihood and deals with continuous outcome variables, such as GPA and salary level. Visualization uses interactive graphs to demonstrate mathematically induced rules and scores, and is far more sophisticated than pie or bar charts. Visualization is used primarily to depict three-dimensional geographic locations of mathematical coordinates [2]. Higher education institutions can use classification, for example, for a comprehensive analysis of student characteristics, or use estimation to predict the likelihood of a variety of outcomes, such as transferability, persistence, retention, and course success.

Data mining tools and algorithms

- Machine learning
- Computer science, heuristics and
- Induction algorithms
- Artificial intelligence
- Emulating human intelligence
- Neural networks
- Biological models and Engineering

Phases of data mining

Data mining is an iterative process that typically involves the following phases:

- Problem definition
- Data exploration
- Data preparation
- Modeling
- Evaluation

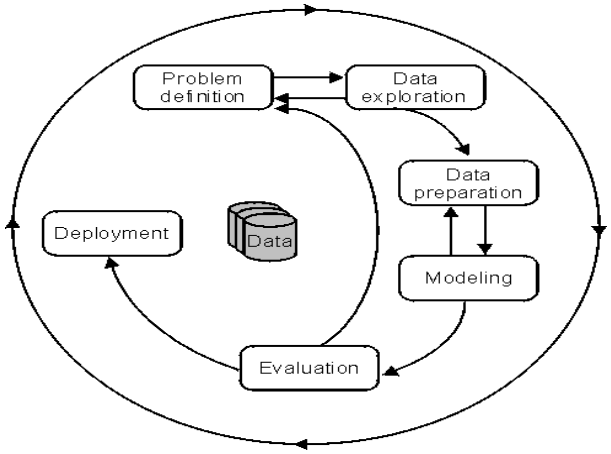


Figure 1 Process of Data Mining

A data mining project starts with the understanding of the problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective [4].

In my project our domain is academic data like a student records, result of colleges of different years, strength of students per year and per department and the experts of that domain are HOD's of departments and principals of the colleges.

In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data. In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed [3].

We select and apply various mining functions because we can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types.

In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required.

Evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When we are finally satisfied with the model, we deployed it.

Tools of Data Collection & Analysis

Various tools are needed for that project some for analyzing data, some for designing, implementation and some developing software tool these are:

- Excel
- Ms access
- J48 algorithm
- Naïve Bayesian Classifier
- WEKA data mining tool
- Tangara data mining tool
- Rapid miner

Advantage of Data Mining In Academics

Data mining tells us following things like:

- Tells us about the weak students.
- Tells us which students are taking more credit hours.
- Subjects which are more interesting to students.
- Type of courses we can offer to attract more students.
- Tells various ways to help weak students.
- Helps in improving the result of schools.
- Helps in predicting the result of students.

2. CONCLUSIONS

The current education system does not involve any prediction about fail or pass percentage based on the performance. The system doesn't deal with dropouts. There is no efficient method to caution the student about the student about the deficiency in attendance. It doesn't identify the weak student and inform the teacher. Another common problem in larger colleges and universities, some students may feel lost in the crowd. Whether they're struggling to find help with coursework, or having difficulty choosing (or getting into) the courses they need, many students are daunted by the task of working through the collegiate bureaucracy. Since the proposed model identifies the weak students, the teachers can provide academic help for them. It also helps the teacher to act before a student drops or plan for recourse allocation with confidence gained from knowing how many students are likely to pass or fail. Proposed system also shows data graphically according to the need or organization which help them to take important decisions. For future work we also use clustering, with the help of clustering we can see the domain and interest of students in particular field [1].

3. REFERENCES

[1]. Hideko Kitahama, "data mining through cluster analysis evaluation on internationalization of universities in japan".

[2]. Bruce l. Golden r. H. Smith School of Business University of Maryland, College park, md 20742 "an example of visualization in data mining"

[3]. Jing Luan, PHD chief planning and research officer, Cabrillo College founder, knowledge discovery laboratories "Data Mining Applications in Higher Education".

[4]. Thulasi, kumarthulasi.kumar@uni.edu, university of northern iowa "theoretical basis for data mining approach to higher education research".

[5]. N.V.Anand Kumar Research Scholar, Department of Computer Science and Engineering, Anna University, Chennai "Improving Academic Performance of Students by Applying Data Mining Technique".

[6]. Han, J. W., Kamber, M., 2006. Data Mining: Concepts and Techniques, 2nd Edition, The Morgan Kaufmann Series in Data Management Systems, Gray, J. Series Editor, Morgan Kaufmann Publishers.

[7]. Luan, J., 2002. Data mining and knowledge management in higher education – potential applications. In Proceedings of AIR Forum, Toronto, Canada.

[8]. D. A. Alhammadi and M. S. Aksoy, "Data Mining in Education – An Experimental Study," International Journal of Computer Applications, vol. 62, no. 15, pp. 31-34, 2013. [8]. B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[9]. Alaa el-Halees (2009) Mining Students Data to Analyze e-learning Behavior: A Case Study.

[10]. U. K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN: 0975-9646, 2011.

[11]. U. Fayadd, Piatesky, G. Shapiro, and P. Smyth, From data mining to knowledge discovery in databases, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–262 56097–6, 1996.

[12]. Berry, M.J.A., & Linoff, G.S. (1997). Data Mining Techniques for Marketing, Sale, and Customer Support, New York: John Wiley & Sons, Inc.

[13]. Chen, S.Y. (1999). Data Mining in Acquiring Association Knowledge, Between Diseases and Medicine Treatments. Unpublished Master's Thesis, National Sun Yat-Sen University, Kaohsiung, Taiwan R.O.C.

[14]. Codd, E.F. (1993). Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate. E.F. Codd and Associates.

[15]. Dai, C.Y. & Sung, J.K. (2005). Data Warehouse of TVC Course in Taiwan, ED-MEDIA 2005, Montreal, Canada, June 27-July 2,2005, AACE.

[16]. Executive Yuan (2002). Promotion Program for Strengthening Digital Content Industry, Executive Yuan of R.O.C.

[17]. Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overviews. Advances in Knowledge Discovery and Data Mining, Menlo Park, CA: AAAI Press.

[18]. Foley, J. (1996). Squeezing more value from data, Information Week, (December 9, 1996): 44.

[19]. Henderson, D. (1996). Tackling an OLAP development project, Computing Canada, 22(19), 24-25.

[20]. Liao, I.Y. (2002). The study on application of data mining for pharmaceutical market in Taiwan, Unpublished Master's Thesis, National Chiao-Tung University, Hsinchu, Taiwan, R.O.C.

[21]. Greenfeld, N. (1996). Data Mining, UNIX Review, 14(5), 9-14.

[22]. Peacock, P R(1998). Data Mining in Marketing: Part 1, Marketing Management, 6(4), 8-18.

[23]. Rob, P & Coronel, C. (2004). Database Systems Design, Implementation & Management, Boston, MA: Course Technology, Thomson Learning, Inc.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 4, 218 - 225, 2016, ISSN:- 2319–8656

A Survey of Packets Scheduling Congestion Control Algorithms in Internet Protocol Storage Area Networks

Joseph Kithinji
Department of Information Technology,
Meru University of Science and Technology
Meru, Kenya.

Abstract: Several approaches have been proposed to empower communication systems with quality of service (QoS) capabilities. In general, their main goal is to coherently support the end-to-end performance needs of applications, based on the establishment of, and agreement on, a set of concepts, policies and mechanisms. The fiber channel is the standard technology used in storage are network communications but it does not have mechanisms for providing QOS guarantees. However the increasing use of transmission control/internet protocol based network storage has introduced the possibility of using already existing techniques and tools to achieve to achieve QOS guarantees in Internet protocol Storage Area Networks. Due to the existence of other competing traffic in internet protocol networks it is necessary to provide storage input/output traffic with guaranteed network bandwidth. This paper discusses the available packet scheduling mechanisms pin pointing their advantages and disadvantages. The main goal is to combine two packets scheduling mechanisms and come up with a hybrid that assure a given storage QOS requirement between a storage client and internet protocol storage.
Keywords: Congestion control, packet scheduling, QOS, prioritization, fiber channel

1. INTRODUCTION

Computer networks were designed mainly to transfer data and email where QOS was effectively implemented by the use of TCP. As storage area networks become popular, many organizations networks have transformed into converged networks in which same infrastructure is shared to ensure all the requested services [20]. Although this convergence offers some advantages like sharing of network media, on the other hand it has come with some disadvantages. One being that it has led to the competition for network resources (buffers of routers), which leads to congestion [18]. Delay in delivering the packets, jitter, loss of packets are consequences of congestion. Since different applications show different sensitivity to these issues. For example, file transfer protocol is not affected by delay and jitter, whereas Storage area networks read requests are very sensitive to packet loss [9]. To solve this problem QOS was introduced to provide better Storage area networks performance.

Due to the increasing need for data storage and economy of scale savings, establishments for storage area networks has been increasing over the last years [19]. In the recent years the storage market has shifted from using expensive fiber channel technology towards TCP/IP based technology. Storage networking is adopting the TCP/IP networking which creates a need for providing QOS in order to offer guaranteed storage performance [14]. This is due to the fact that in TCP/IP networks there are other competing traffic for network bandwidth.

Guaranteed storage performance is an essential requirement for applications using it, and it's important for network designers to ensure that storage performance meets the requirements of the applications utilizing it. In a storage area network**,** a single host request may flood the resources of a storage pool causing poor performance of all hosts utilizing that particular pool [14]. Hence, the performance of a given host utilizing a shared pool resource is unpredictable by the nature of resource sharing. To address this problem a mechanism of providing QOS based on some policy is required. Storage service level agreements provide for predictability in service delivery which is not effective due to the absence of QOS mechanisms in storage devices [13]. However when we utilize TCP/IP as transport mechanisms, packet scheduling mechanisms which have been used and studied well are available.

Internet small computer interface is the technology used for implementing the IP-SAN [10]. The Internet Small Computer Systems Interface (iSCSI) is a TCP/IP – based protocol for establishing and managing connections between IP-based storage devices, hosts and clients. It defines the rules and processes to transmit and receive block storage applications over TCP/IP networks by encapsulating SCSI commands into TCP and transporting them over the network via IP. It is a protocol for new generation of data storage systems that natively use TCP/IP [14]. Since internet small computer system interface uses TCP for transportation, it is possible to use the available packet scheduling algorithms tools for the purpose of throttling traffic destined to storage devices.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 details the existing packet scheduling algorithms. Section 4 presents hybrid queues. Section 5 presents our discussion while section 6 concludes the paper.

2. BACKGROUND

Online data storage doubles every nine months due to the increasing demand for networked information services [13]. As a result network storage architectures have developed from network attached storage, to storage area network and most recently storage over IP (IP SAN). Internet SCSI (internet small computer system interface) is the technology used for implementing the IP-SAN. Providing QOS to Storage area networks has been a challenge which has led to the design of many approaches such as Stonehenge, cello, facode, triage, argon, chameleon and Aqua[16]. Despite all these research specification regarding QOS in Storage area networks utilizing TCP/IP have not been exhausted.HPLab storage systems department has been researching on how to provide Qos(based on response time and bandwidth)[16]. In this case bandwidth is to be allocated on demand.

Storage area networks technologies such as fiber channel and gigabit Ethernet have enabled storage systems to be maintained as storage pools, reducing the total cost of ownership[13,16]. Few protocols have been developed to support storage area network environment. Fiber channel protocol was developed for fiber channel based storage area networks. Internet small computer system interface was developed for internet protocol based storage area networks. A main advantage of internet small computer system interface is that it can operate on standard Ethernet; that is it can be able to exploit existing features and tools that have been developed for internet protocol networks [18, 19]. Thus this paper focuses on the storage environment using internet protocol where storage devices are attached to internet protocol networks and storage clients communicate with the storage devices via the internet small computer system interface.

A different storage client may require a different storage service called storage quality of service; that is each client requires receiving a guaranteed storage service independently of the status of the input/output services in other storage clients [8, 9, 11]. The use of internet protocol in storage area networks avails the commonly used QOS implementation techniques which can be applied in storage area networks.

The combination of well-known and trustworthy throttling mechanisms and an extended knowledge about storage systems internals makes an appealing pragmatic and non-intrusive approach to the problem of QOS in storage systems [17]. Instead of building new scheduling algorithms for storage devices, this paper suggests the utilization of previously known and trusted tools to implement QOS in storage area networks [6]. Since fiber channel does not provide prioritization, the use of transmission control/internet protocol as a transport mechanism makes the commonly used packet scheduling algorithms available.

The market for internet small computer system interface storage devices is growing making it an interesting target for quality of service research [8]. The integration of the well-known TCP/IP throttling mechanisms and storage systems internals provide a good approach to solving the problem of QOS in storage systems [7]. Adoption of existing systems would be more appropriate instead of introducing new algorithms which are bound to cause uncertainty and overhead.

3. PACKET SCHEDULING ALGORITHMS

3.1 FIFO Queuing

FIFO is the most widely used queuing discipline because of its easy configurations [2]. Packets belonging to different flows pick up in the FIFO queue and processed in the order of arrival.FIFO belongs to the unconscious group which treat packets as they are. Packets from all input flows are queued into a memory stack, then they are dequeued in the order of arrival one by one onto the output link. Since FIFO does not perform any reorganization of the queue, there is no schedule overhead experienced by packets [1]. This means turnaround time, waiting time and response FIFO time are low. However due to the lack of prioritization, FIFO systems may have trouble meeting deadlines.

On the other hand lack of prioritization ensures that every process will eventually complete its transmission without the risk of starvation [1]. All packets are placed in a single queue and are treated equally [4]. Packets are transmitted as bandwidth becomes available. Packets are treated accordingly to their arriving order. However since the queue buffer is finite, any packets that arrives after the buffer if full it is dropped without regard to which flow the packet belongs to or how essential a packet is [5]. This concept is known as the trail drop concept.

FIFO works well in links that are not heavily congested .Since works on first come first serve basis, if a node initiates a large file transfer, it can consume all the bandwidth link to the disadvantage of other traffic [18]. This phenomena is known as packet trains since the source sends a train of packets to its destination and packets from other hosts get caught behind the train. FIFO is efficient for large links that have little delay and minimal congestion [5]. To optimize QOS metrics such as buffer requirements and queue delay, FIFO uses traffic shaping techniques [6]. FIFO also employs AQM mechanisms to ensure fairness among flows.

3.2 Priority Queuing

Packets are assigned to classes which are associated with certain priority value. Packets with high priority are processed first. Priority queue is able to differentiate traffic hence reducing delay of important traffic. On the other hand if there is continuous flow of high priority traffic, low priority will be starved [6]. In basic implementations of priority queuing, it consists of four priority queues where packets are handled using FIFO. Packets belonging to the highest priority queue are serviced first. A packet scheduler is used to check for existence of packets in the highest priority queue after the current packet is processed. Any packets that arrive in the high priority queue are processed immediately [9]. The main advantage of priority queuing is that it is able to ensure highest priority to storage area networks(especially for read requests) but also lead to infinite delay for packets belonging to lower priority queues.

Priority queuing technique is suitable for situations where mission critical traffic needs preference [8]. Priority queue provides a smooth transition of important traffic (packets), through the network, using management at all intermediate points. Priority queuing classifies traffic with priority labels low, normal, medium and high. Packets which have not been assigned to a class automatically fall into the normal waiting queue. Data belonging to the high priority queue is handled first followed by that belonging to low priority queues. Priority queue currently uses static configuration and because of this it is not able to automatically adjust to the changing requirements in the network [14]. All incoming queues are assigned to a given network interface with each queue having a priority level.

When queues are being sent out the interface, they are scanned for packets in descending order of priority. The high priority queue is scanned first followed by the lower priority queues. The packet at the head of the highest queue is chosen for transmission [19]. This process is repeated every time a packet is to be sent [15]. Priority queuing has four traffic priorities; high, normal, medium and low (comparative study of different queue disciplines). Priority queuing is useful in environments where we want to make sure that mission critical traffic gets priority treatment. However priority queuing has a weakness in that it does not automatically adapt to changing network requirements due to static configurations [6]. Although priority queuing is simple, its implementation, it can cause queuing delay and increased jitter on the lower traffic.

3.3 Class based queuing

Priority queue is able to differentiate traffic hence reducing delay of important traffic. On the other hand if there is continuous flow of high priority traffic, low priority will be starved [6]. In basic implementations of priority queuing, it consists of four priority queues where packets are handled using FIFO. Packets belonging to the highest priority queue are serviced first. A packet scheduler is used to check for existence of packets in the highest priority queue after the current packet is processed. Any packets that arrive in the high priority queue are processed immediately [9]. The main advantage of priority queuing is that it is able to ensure highest priority to storage area networks(especially for read requests) but also lead to infinite delay for packets belonging to lower priority queues.

Priority queuing technique is suitable for situations where mission critical traffic needs preference [8]. Priority queue provides a smooth transition of important traffic (packets), through the network, using management at all intermediate points. Priority queuing classifies traffic with priority labels low, normal, medium and high. Packets which have not been assigned to a class automatically fall into the normal waiting queue. Data belonging to the high priority queue is handled first followed by that belonging to low priority queues. Priority queue currently uses static configuration and because of this it is not able to automatically adjust to the changing requirements in the network [14]. All incoming queues are assigned to a given network interface with each queue having a priority level.

When queues are being sent out the interface, they are scanned for packets in descending order of priority. The high priority queue is scanned first followed by the lower priority queues. The packet at the head of the highest queue is chosen for transmission [19]. This process is repeated every time a packet is to be sent [15]. Priority queuing has four traffic priorities; high, normal, medium and low (comparative study of different queue disciplines). Priority queuing is useful in environments where we want to make sure that mission critical traffic gets priority treatment. However priority queuing has a weakness in that it does not automatically adapt to changing network requirements due to static configurations [6]. Although priority queuing is simple, its implementation, it can cause queuing delay and increased jitter on the lower traffic.

3.4 Fair queuing

The fair queuing is a scheduling mechanism that classifies packets and processes these packets based on a service level agreement. Fair queuing uses a round robin algorithm to allocate bandwidth where every flow has an equal chance [6, 10]. The round robin algorithm ensures that flow from one class does not starve other classes off the bandwidth. The main advantage of priority queuing is that in a situation where there is congestion in a particular class, other classes are not affected and therefore the overall network performance is not affected. The downfall of priority queuing as a scheduling mechanism is that it does put into consideration the packet length. This means if a particular class has big flows, then the class may use more bandwidth and therefore take longer to be served [8]. However fair queuing is considered to be best suited in sharing bandwidth among different classes with the same bandwidth requirements.

3.5 Weighted Fair Queuing

In weighted fair queuing, incoming packets are grouped into classes and admitted to different queues[7]. Then these queues are assigned priority based on their weights, with high weights corresponding to high priority. Packets are then processed in a round robin manner with the number of packets selected from each queue based on the corresponding weight. For example in a case where we have weights 3,2 and 1,this would mean that three packets are processed from the first queue ,two from the second queue and one from the third queue. If the bandwidth manager does not impose priority on the classes, all weights can be equal. In this way we have fair queuing with priority [6]. All configurations in weighted fair queuing are automated with no room for tuning possibilities.

Queues from flows are grouped into a maximum of 256. The weighted fair queuing then uses the following notations:

SN=previous-SN (Weight*new packet length)…1

Weight =32384/ (IP-precedence+1)…..2

Where SN is the completion time [9].

Weighted fair queuing is mostly suitable in environments where it is desirable to provide a constant response time for the demanding users or applications without adding an excessive bandwidth. Weighted fair queuing implements bitwise fairness, which allows a queue to be served based on the number of bytes [9]. Weighted fair queuing ensures no traffic is starved off bandwidth. In this way low-level traffic can smoothly travel through the network. This increases service efficiency since an equal number of low-level and high level packets are transmitted. Weighted fair queuing can also automatically adapt to the changing network conditions. The weights are calculated from IP priority bits where values 0 to 5 are used (6 and 7 are reserved) and the weighted fair queuing algorithm calculates how many additional services must be provided for every queue [5]. Weighted fair queuing reduces the round trip delay which makes it perform better than TCP and in the process reducing congestion and speeding up slow connections.

Weighted fair queuing results in predictable behavior over the entire route while the response time for each active flow can be reduced by a multiple factor [2]. The weights are used also to determine how much bandwidth each flow is allocate relative to others, the maximum length of a queue is defined by the length limit [13]. Weighted fair queuing disciplines sorts packets in weighted order of arrival of the last bit to determine the transmission order. Transmission order bits can be used to identify weights. Weighted fair queuing is aware of packet sizes and can support variable sized packets, so that flows with large packets are not allocated more bandwidth than the queues with smaller packets [5]. Flows are grouped into those requiring huge amounts of bandwidth [6]. The goal is to always have bandwidth available for the low throughput flows to split the rest proportionally to their weights.

Since weighted fair queuing is derived from fair queuing, if N data flows are currently active with weights w1, w2…wn,

data flow number i then average data rate can be achieved using the equation below [13]

Average data rate=rwi/w1+w2+……wn.

By regulating the weights, automatically, weighted fair queuing can be used to achieve guaranteed data rate. Each flow is allocated with different bandwidth percentage hence preventing monopolization of the bandwidth by some flows.

3.6 Class Based Weighted Fair Queuing

Class based weighted fair queuing is based on the idea of weighted fair queuing, the only difference is that in class based weighted fair queuing, traffic flows are grouped in classes. The other difference is that classification is done manually in class based weighted fair queuing unlike in weighted fair queuing where configurations are done automatically [20]. This ensures flexibility in allocating a minimum bandwidth amount on the fair queuing basis as well as on the basis of administrator defined classes [2, 3]. If a traffic flow is not attached to any configured classes, it can use only the remaining link bandwidth which is not associated with any class. Each class is allocated a guarantee amount of bandwidth. Class based weighted fair queuing is used in situations where more low-priority flows could overflow the high priority data stream. However low latency queue can be marked so that actual high priority queue is differentiated. Traffic belonging to low latency queuing will be serviced before all other traffic placed in other classes and at the same time the necessary amount of bandwidth will be guaranteed [4]. When a class does not use all the guaranteed bandwidth, it can be shared among other classes.

3.7 Custom Queuing(CQ)

In CQ flows are categorized into 16 FIFO queues with a defined buffer length. Each of the FIFO queues is then assigned a suitable percentage of the total bandwidth. Scheduling of the queues in the output interface is done in round robin [2]. Although CQ can be able to eliminate the infinite delay experienced in priority queuing, it cannot implement priority for storage area networks. However a fine tuning of row lengths can help to reach acceptable results [9]. CQ guarantees mission critical flows a certain percentage of the whole bandwidth while assuring other traffic will get predictable through put [2]. For example if we have 16 queues, queue 0 is configured as a special queue called system queue which is used to handle keep alive and control packets that are considered as high priority packets. Queues 1 to 15 are used to carry user defined traffic. This means user traffic cannot pass through queue 0.

Traffic can be classified based on input interface access control lists, application types and packet sizes. Queues are then served in a round robin manner until a byte counter limit threshold is met. Once this threshold is met, the frame from the next queue are serviced [4]. In CQ routers service each queue sequentially transmitting a configurable percentage of traffic on each queue before moving to the next one. CQ routers determine how many bytes should be transmitted for each queue based on the interface speed and the configured percentages. When a particular queue is being processed, packets are sent until the number of sent bytes exceeds the byte count, or until the queue is empty [5].In this way all the traffic is processed.

3.8 Modified Weighted Round Robin

MWRR uses variable –sized packets to determine which queue to be served. To calculate the variable size, it uses a deficit counter variable to initialize to each queues weight. Before a queue is serviced, its deficit counter is initialized to the queues weight. A packet is scheduled for transmission if the deficit counter is greater than zero.

High priority packets are allowed to cut front of the line. The processed number of packets is equal to the normalized weight over the mean packet size [8, 11]. MWRR queuing discipline serves packets at the head of every non empty queue whose modified counter is greater than the size of the packet at the head of the queue.

3.9 Deficit Weighted Round Robin(DWRR)

DWRR was proposed by M.Shreedher and Varghese in 1995 [20]. DWRR handles packets of variable size without any consideration of the mean size. The number of the packet size is subtracted from the packet length and packets that exceed the packet number are held back until the next visit of the scheduler. DWRR uses scholachastic fair queuing to assign data flows to queues [17]. Queues are served in a round robin with a quantum of service attached to each queue [16]. This implies that if a queue is unable to send packets due the size, the remainder from the previous quantum is added to the quantum for the next round.

Queues not serviced in a round are compensated in the next round. However once a flow is serviced it must wait for N-1 other flows to be serviced before it is serviced again. During each round, a flow transmits its entire quantum data once as a result DRR has poor delay [7]. Each queue is associated with a quantum and a deficit counter. Quantum represents the number of bytes that each queue can send on its turn [3]. The deficit counter variable is used to keep track of the credit each queue possesses for sending traffic and is initialized to zero.

3.10 Modified Deficit Round Robin(MDRR)

In MDRR when a queue is served a fixed amount of data in bytes is dequeued. The algorithm then services the next queue. If the amount of data dequeued exceeds the value configured, in the next round less data will be dequeued to compensate for the excess data that was previously dequeued. As a result, the average amount of data dequeued per queue will be close to the configured value [3, 4]. Queues in MDRR are defined using two variables; a quantum value (number of bytes served in each round) and a deficit counter (used to track how many bytes a queue has transmitted in each round).

Packets are served only when their deficit counter is greater than zero. Each packet served decreases the deficit counter by a value equal to its length in bytes [6]. Queues with deficit counters as zero or negative are not served. In each new round the deficit counter of each non-empty queue is increased by its counter.

4. HYBRID WAITING QUEUES

Because different queuing mechanisms have different advantages, the idea is to combine different queuing mechanisms and join their positive but also negative properties into new hybrid queuing method. The main

disadvantage of hybrid queuing is the duplication of the memory of the mechanism which forms a queue [19]. This is due to the fact that every memory element and its size involve certain latency or delay for traffic which goes through these interfaces. The higher the number of these interfaces the bigger the delay which is detrimental to applications such as Storage area networks read requests [3, 5, 8]. This is why we have to make a compromise between the number, size and length of the intermediate buffers to avoid excessive data spillage (or data loss). When a buffer is too small and to also avoid scenarios where the buffers are too big, and are increasing the delay. This aspect is called the jitter effect.

4.1 The custom queuing –class based weighted fair queuing hybrid waiting queue

Combine CQ and class based weighted fair queuing. In the first phase CQ is used to allocate bandwidth among all active applications to avoid congestion. Once packets are sent to the output CQ interface they arrive to the class based weighted fair queuing input interface [20]. Class based weighted fair queuing arranges traffic into classes defined by a class-based weighted fair queuing algorithmically the advantages of the class based weighted fair queuing are retained[3]. With this method we reduce the delays within the network, which is not the case with ordinary CQ scheme**.**

4.2 Priority queuing-class based weighted fair queuing hybrid waiting queue

In the first step priority queuing is used to arrange flows into waiting queues according to the priorities set in individual packets TOS fields[6]. The output interface of priority queuing algorithm first serves the highest priority data streams and then all other lower ranking queues. As packets leave the priority queuing interface they have already been assigned bandwidth as configured by the network administrator. This way the packets at the class based weighted fair queuing mechanism output interface do not need to fight for bandwidth as it is guaranteed in advance [5]. This accelerates the transfer of high priority flows and such flows become independent of all other lower priority flows.

The priority queuing-class based weighted fair queuing is closely related to low latency queues. The low latency queuing mechanism allows a class that is served as a strict priority queue. Traffic in such class will be served before all other traffic in the remaining classes. Bandwidth amount is also guaranteed in this case [3]. This implies that all traffic which is above the level of bandwidth reservation is simply discarded

4.3 Weighted fair queuing-class based weighted fair queuing hybrid waiting queue.

Weighted fair queuing is implemented in the first step to ensure fairness for all applications since internal weighted fair queuing are emptied by principle of fairness. Weighted fair queuing ensures that undisturbed flow throughput for all active applications. In the next step the class based weighted fair queuing puts packets into admin defined classes. This way every application at the first stage gets a fair treatment and in the second phase high priority applications gets its own classes with the pre-reserved bandwidth. The rest of the bandwidth is left for all other active applications [6, 7]. The fairness and fluidity movement apply for all active applications.

Weighted fair queuing is effective for operating with IP priority settings such as resource reservation protocol. Weighted fair queuing is also capable of managing round trip delay problems. This is the main reason for combining the weighted fair queuing and class based weighted fair queuing [20]. The weighted fair queuing-class based weighted fair queuing is at the same time capable of accelerating slow features and removing congestion in the network. The results become more predictable over the whole routing path, while Ethernet delays can be greatly decreased compared to CQ, priority queuing and weighted fair queuing [3]. The weighted fair queuing and the class based weighted fair queuing combination can represent the best solution for reducing the Ethernet delay.

5. DISCUSSION

In this paper we have looked at packet scheduling algorithms that can be used to reduce congestion in storage area networks. We have looked at the strengths and weaknesses of each algorithm. First we looked at FIFO which is considered a simple mechanism since when there is no congestion packets do not experience any delays. On the other hand if there is congestion packets will experience delays due as a result of queuing delay. Therefore FIFO cannot be used to provide quality of service guarantees. This makes FIFO a very weak scheduling technique due to the fact that internet traffic is always bursty which in all cases leads to congestion. This weakness of FIFO makes it no to be implemented alone.

We have also looked at priority queuing which has got an advantage in that in a situation where there is congestion in a particular class, other classes are not affected and therefore the overall network performance is not affected. This ensures that a particular class misusing its bandwidth does not affect the other classes.

The downfall of priority queuing as a scheduling mechanism is that it does put into consideration the packet length. This means if a particular class has big flows, then the class may use more bandwidth and therefore take longer to be served. This may end up starving other classes and does not allow bandwidth sharing. Priority Queuing is most appropriate where we have a fixed number of queues requiring different priorities. But due to the fact that the high priority traffic is transmitted first, priority queuing cannot be used to offer end to end service guarantees. However priority queuing can be combined with leaky bucket algorithm to ensure that high priority queues do not monopolize the link. However fair queuing is considered to be best suited in sharing bandwidth among different classes with the same bandwidth requirements. This is because with fair queuing every class gets a guaranteed fair share of the bandwidth with allowance for sharing if not in use. Fair queuing ensures that all traffic has fair access to network resources. This prevents bandwidth starvation for less aggressive traffic.

Weighted fair queuing attaches weights to every flow which determines the amount of bandwidth associated with a particular flow hence guaranteeing a constant rate of bandwidth. Configurations in weighted fair queuing are do automatically which makes possible to provide QOS guarantees since it is able to adapt to changing network conditions. Class based weighted fair queuing improves on the weighted queuing by classifying flows so as to offer differentiated service. Configurations are also done manually which provides more flexibility in assigning bandwidth to traffic flows. Flexibility in configurations allows for network administrators to vary the configurations now and then which leads to a more customized bandwidth allocation for a particular organization.

Since different queuing mechanisms have got different advantages, they can be combined into hybrid queues to offer better quality of service guarantees. CQ can be combined with class based weighted fair queuing which would result in reduction delays associated with CQ. Priority queuing and class based weighted fair queuing can be combined to provide bandwidth guarantees not available in priority queuing. The combination of queuing disciplines in congestion control may result in improved results however there is added overhead due to the processing required by the particular combined mechanisms.

6. CONCLUSIONS

In this paper we have discussed packet scheduling techniques that can be used to achieve quality of service in storage area networks. We have discussed each technique separately in order to get insight on its strengths and weaknesses. From our discussion it is evident that no single technique can achieve quality of service guarantees in storage area networks. Hybrid approaches may result in better quality of service guarantees but, results in jitter affect which is caused by duplication of the memory mechanisms where the queues are formed. This is because every memory element and its size involve a certain latency for traffic which goes through that interface.

In future research we aim at exploring techniques of reducing the jitter effect caused by combining scheduling mechanisms.

7. ACKNOWLEDGMENTS

My thanks goes to Professor Muchiri Muketha for his advice in the development of the paper.

Table 1.Comparative analysis of packet scheduling algorithms in providing QOS

MEASURE OF QOS	**FIFO**	**priority**	**fair queuing**	**weighted fair queuing**	**custom queuing**	**Modified Weighted Round Robin**	**Deficit Weighted Round Robin**	**The custom queuing – class based weighted fair queuing hybrid waiting queue**	**Prio base que wait**
starvation	High	High	Low	Low	Very low	Medium	Medium	Low	Low
Packet loss	High	High	Very low	Very low	Low	Low	High	Medium	Med
Bandwidth guarantee	Low	Low	Medium	High	High	Low	Low	High	High
delay	High	Very high	Medium	Low	Low	Medium	Medium	Low	Low
Jitter	High	High	Low	Low	Low	Low	Low	Very low	Very

8.REFERENCES

[1] Harpreet, K, Gurpal & S, Fatehgarh. (2011)" Wimax Networks Implementation and Evaluation of Scheduling Algorithms in Point-to-Multipoint Mode", IJCST Vol. 2, Issue 3, India.

[2] Ahmad, K & Bahauddin, Z (2011)."VoIP Performance Over different service Classes under Various Scheduling Techniques", Australian Journal of Basic and Applied Sciences, 5(11): 1416-1422-CC, ISSN 1991-8178.Pakistan.

[3] Sasa K, Amor C, Joze M & Zarko C.(2012)","Influences of Classical and Hybrid Queuing Mechanisms on VoIP's QoS Properties" .University of Maribor.

[4]Rajeev,S, Sukhjit & S, Sumeet.(2015),"International Journal of Advanced Research in Computer and Communication Engineering "Vol. 4, Issue 3,india.

[5] Sarhan M. Musa, Mahamadou T, Matthew N. O. ,Pamela O & Roy G. P.(2013)."A Comparative Study of Different Queuing Scheduling Disciplines", Journal of Engineering Research and Applications Vol. 3, Issue 6, pp.1587-1591, Malaysia.

[6] Nidhi, M & Anil, K.(2013)."Active Scheduling (Queuing) Algorithms in Congestion Management: A Review ", International Journal of Digital Application & Contemporary research Volume 1, Issue 8, India.

[7] Ahmed K. (2011)"VOIP performance over different service classes under various scheduling techniques", Australian journal of basic and applied sciences (11):1416-1422, Pakistan.

[8] Sadafale, K, Barahate & Prashant, J. (2010),"Improvement and analysis of voice data traffic in VOIP", International conference and workshop on emerging trends in technology (IC WET).ACM, New York.

[9] Szabolcs, S. (2013)."Analysis of the algorithms for congestion management in computer networks", Carpathian Journal of Electronic and Computer Engineering 6/1 3-7 3 ISSN 1844 – 9689, Hungary.

[10]Sulbha, S, Hemke, V, Gawande, D & Gautum,L.(2013)."ISCSI-the future of storage Network ", international journal of application or innovation in engineering and management Volume 2, issue 4,ISSN 2319-4847,Australia.

[11] Rajesh K & Vinay.(2012)."Performance Evaluation of Scheduling Algorithms in WLAN Network with CBR Application using Qualnet", International Journal of Electrical, Electronics and Computer Engineering 1(1): pp1-5 ISSN No: 2277-2626,India.

[12] Dhaini, R,Assi, C.M,& Shami,A.(2008),"Dynamic bandwidth allocation schemes in hybrid TDM/WDM passive optical networks",IEEE Consumer Communications and Networking Conference.Vol.1,no.7,Germany.

[13] Prasad, R, Murray, M, Dovrolis, C, & Claffy, K. (2011) Bandwidth Estimation: Metrics, Measurement Techniques, and Tools, IEEE.Vol. 17, no. 6(May), pp. 27-35,Scotland.

[14] Devajit, M, Majidul A & Utpal, J.B. (2013)A study of Bandwidth Management in Computer Networks, International Journal of Innovative Technology and Exploring Engineering (IJITEE) .Vol.2, no.2,pp 20-30,Australia.

[15] Farhangi, S & Rostami, S.(2013),"A Comparative Study Between Combination of PQ and MWRR Queuing Techniques in Ip Network Based on OPNET", Middle-East Journal of Scientific Research 13 (8): 1051-1056, 2013 ISSN 1990-9233, IDOSI Publications,Turkey.

[16] Borgeengen, J, Haugerud, H.(2010),"Using traffic shaping to achieve ISCSI service predictability", 24th Large installation system administration conference.USENIX association, Norway.

[17]Van der Stok,P, D. Jarnikov,D,Kozlov, S,van Hartskamp,M & Lukkien,J.J(2009),"Hierarchical Resource Allocation for Robust In-Home Video Streaming", Journal of Systems and Software.Vol. 80, no. 7, pp. 951–961,Ireland.

[18]Van Rensburg, J.R, Veldsman, A, & Jenkins, M. (2008).From technologists to social enterprise Developers, ICT for development practitioners in Southern Africa.Vol.14, no.1, pp 76–89,Australia.

[19]Xu, X, Liang, D, Jiang,H & Lin,X.(2009)Dynamic Bandwidth Allocation in Fixed BWA Systems,Proceedings of the International Conference on Communication Technology (ICCT), vol.2,no.6, pp.1000-1003,Shangai.

[20]Yi-Nung, L, Meng-Che, C, & Shao-Yi, C.(2009)Bandwidth and local memory reduction of video encoders using Bit Plane Partitioning Memory Management,IEEE International Symposium on .Vol.14, no.8, pp.766-769, 24-27,Shangai.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 4, 226 - 233, 2016, ISSN:- 2319–8656

A Review of Intrusion Alerts Correlation Frameworks

Joseph Mbugua Chahira
GarissaUniversityCollege,
Garissa- Kenya

Jane Kinanu Kiruki, Chuka University,
Chuka- Kenya

Peter Kiprono Kemei
Egerton University,
Nakuru- Kenya

Abstract : The advancement of modern computers, networks and internet has led to the widespread adoption and application of Information Communication Technology in modern organizations. As a result, large amount of information is generated, processed and distributed through digital devices. On the other side, digital crimes have increased in number and sophistication and they compromise the organization's critical information infrastructure affecting the confidentiality, integrity and availability of its information resources. In order to detect these malicious activities, organizations deploys multiple Network Intrusion Detection Systems (NIDSs) in their corporate networks. They generate huge amount of low quality alerts and in different formats when an attack has already taken place. Thus Alert and event correlation is required to preprocess, analyze and correlate the alerts produced by one or more network intrusion detection systems and events generated from different systems and security tools to provide a more succinct and high-level view of occurring or attempted intrusions. This work will review current alert correlation systems in terms of approaches and propose design consideration for an efficient alert correlation technique. We conclude by highlighting the opportunity to include attack prediction component in a real time multiple sensors environment.
Key words alert correlation, Intrusion Detection Systems, Attacks prediction, Attack strategy, Network security.

1.0 INTRODUCTION

The modern enterprise relies on network enabled applications, distributed rather than centralized computing resources and internet access to every networked device to conduct and improve business transactions. As a result large amount of information is generated, processed and distributed electronically (Sommer, 2009). Against the background of these accelerating technological changes and disruptive business models, enterprises with online presence are at a high risk of cyber-attacks and threats. Cyber crimes have increased in frequency and their degree of sophistication has also advanced ((Healy, 2008, Alharbi, 2011). Finding the most effective way to secure information systems, networks and sensitive data based on the current trends is a challenging task experienced by many organisation.

Information Assurance and Security (IAS) is a crucial component in the corporate environment to ensure legitimate access to critical information resources and prevent illegal alterations, protect the trustworthiness of information and maintain secrecy of sensitive information against unauthorized access. The organizations have implemented various Intrusion Detection and Prevention Systems (IDPS)on protecting the information on a secured network. However, they are unable to provide bullet proof protection that copes with the large amounts of information to be protected, the advancement of cyber threats and attacks and also the current manual way of analyzing alerts which istedious, time consuming and error prone for the security analyst to manage and verify true alerts (Maheyzah at el, 2015; wang at el, 2005).

In order to optimally discover the complete relationships among alerts from multiple sources a more practical and efficient Alert Correlation system is required (Maheyzah at el, 2015; Rahayu at el, 2009). This will address the problem of huge volumes of low quality alerts, computational requirements and increase the accuracy of generated output. Indeed, the complete discovery can reveal the behavior of the attacker and predict the next steps of action in a proactive step to assist response systems react before the network is compromised.

The following section presents the overview of the main correlation approaches and proposed frameworks that falls in these areas of intrusion alert correlation approaches. Section Three provides a summary of these approaches in terms of the strengths and limitations and design Considerations for an effective alert correlation technique. Lastly, we conclude the paper and present potential future work.

2.0 ALERT CORRELATIONS SYSTEMS

The research on alerts clustering and correlation techniques have been carried out for several years to provide a global view of attacker's behavior by analyzing low-level alerts produced by the IDS sensors and other security systems. The main objective of alerts correlation is to build an automated abstract modeling of alerts by reducing the number of meta alerts generated from alert aggregation process (Fatma at el,2013; Bateni et al., 2012). The correlation system achieves this by identifying and suppressing false alerts, grouping alerts that refer to the same incident together, constructing attack scenarios, prioritizing the alerts, attack detection and prediction (Maheyzah at el, 2015; Rahayu at el, 2008). The main approaches of alert correlation techniques can be divided into: Similarities of Alert Attributes technique, Predefined Attack Scenario, Prerequisites and Consequences of Attacks, and data mining and expert based

2.1 Similarities of Alert Attributes

This approach focuses on improving the quality of alerts reduction by comparing an alert to all alert threads that have similar attributes or features (such as Source IP address, source port number, destination IP address, destination port number, and time stamps). A correlation score is calculated between these alerts and then correlates alerts with a high degree of feature similarity if match or a new thread is created if none is match depending on the score. This method is simple and easy to implement, but is unable to detect complex attacks due to its reliance only on expert knowledge to determine the similarity degree between attack classes (Karunasekera at el, 2010, rahayu at el , 2008). In addition, it fails to discover the causal connection between alerts when alerts with different attributes have been induced in a single attack. In this case, not all the attacks can be detected.

Collection mechanism and reduction of IDS alert framework (CMRAF) (Al-Saedi at el, 2012) was proposed to remove the duplicates IDS alerts and reduce the number of false alerts. They use information gain ratio algorithms to extract the similarities between set of alerts and provide the highest weight to the most effective features based on the class of alerts belonging to the algorithm.

Alert correlation using a novel clustering approach, (Mohamed et al. 2012), applied an incremental clustering approach to reduce the amount of alerts generated by IDS. Three attributes, destination IO, signature-id, and timestamp had been extracted and hashed by using MD5. The hash value from the next input tuple is checked against hash value of the existing clusters. The hashing technique is used to speed up the comparison in checking the similarities of alert attributes.

An improved framework for intrusion alert correlation, (Elshoush at el, 2012), divided alert correlation into 10 main components and contained them in the Data Normalization Unit, Filter-based Correlation Unit and Data Reduction Unit. Similar alerts are fused based on seven extracted features, namely EventID, timesec, SrcIPAddress, DestPort, DestIPAddress, OrigEventName, and SrcPort in order to remove duplicate alerts created by the independent detection of the same attack by different sensors.

Valdes et al. (2008), proposed a probabilistic-based approach to correlate and aggregate security alerts by measuring and evaluating the similarities of alert attributes. They use a similarity metric to fuse alerts into meta-alerts to provide a higher-level view of the security state of the system. Alert aggregation and scenario construction are conducted by enhancing or relaxing the similarity requirements in some attribute fields. But similarity calculations the only way for them to aggregate the alerts. They have to compare all the alert pairs and have to determine lot of thresholds with expert knowledge which lead to their huge volume of computing workload.

2.2 Predefined Attack Scenario

This technique utilizes the fact that intrusions often require several actions to take place in order to succeed. Every attack scenario has corresponding steps required for the successfulness of the attack. Low- level alerts from IDS are compared against the pre-defined attack scenario before the alerts can be correlated. It is restricted to known attack and misuse detection only and specified by human users or learned through training datasets.

The main advantage for this method is that it is able to accurately detect well-documented attacks derived from the libraries. But if it is a novel attack, the method will fail to detect the intrusion (Osman at el, 2010, Alsaidi, at el 2012, Benferhat at el, 2012) Limitations of this approach are the need of more complete and comprehensive scenario libraries; time and cost to build and maintain them are the main concerns. In-depth knowledge on various attack scenarios is required to manually define the attack conditions. Thus, it may not be practical for complex and large-scale networks.

An alert fusion model inspired by artificial immune system, (Mahboubian at el, .2012) which is an aggregating method inspired by Danger Theory to aggregate the generated alerts based on the prediction of attack scenarios. They categorized network attacks into three general groups which are One-to-One, Many-to-One, and One-to-Many. Each alert will be grouped and a priority value will be given. Then each group is checked with predefined rules for the possibility of raising the danger alarm by using the Danger Theory.

Automatic attack scenario discovering based on a new alert correlation method (Ebrahimi at el, 2012) introduced a method to automatically extract multi-step attack scenarios. They arrived alert had been determined as to which alerts scenario it belongs to and inserted in an alert tree. Sub scenarios in each scenario and meta-alerts are extracted. Finally, the multi-step attack graph is constructed for each attack scenario from the produced meta-alerts.

A novel algorithm for the alert correlation generated by signature-based network IDS (NIDS) had been presented by (Marchetti at el, 2012). The proposed algorithm called pseudo Bayesian alert correlation is based on Bayes's theorem of conditional probability. This algorithm aims to identify and highlight groups of intrusion alerts based on their detection time and on the past alert history generated by same NIDS. In this case, the previous alert history was analysed periodically while recent intrusion alerts are received from the NIDS and analysed as soon as they are generated.

2.3 Prerequisites and Consequences of Attacks

Most attacks today are not isolated but related to each other as different stages of attack sequences with the earlier attacks paving way for consequent attacks. In order to ensure the attack is successful, the prerequisite of an attack is the necessary condition (Ning at el, 2004). In this approach, a set of detailed criteria is used to learn the causal relationship between alerts and the weights of such relations. The main benefit to network analysts when using this method is that they do not have to specify all the possible scenarios but they are still able to detect unknown attacks. Nonetheless, it is expensive (in terms of human expertise) to build a complete attack database which consists of every attack action with its pre- and post-conditions (Karunasekera at el, 2010, fatma at el, 2013). Similar to predefined attack scenario approach discussed in

the previous section, this approach may not be practical in production networks due to the complexity of the design and user behaviour.

Bayesian network-based alert correlation, (Anbarestani at el, 2012) discovers attack strategies without the need for expert knowledge. The approach extracted attack scenarios using classification by taking into account the sequence of actions. It then leverages upon historical data from log sources and classifies them based on observed intrusion objective as class variables. The possible attack scenarios constructed from hyper alerts sequences are examined and the most plausible strategies for constructing a cooperative attack are extracted.

Zhaowen at el (2010), used an on-line prerequisite-consequence-based correlation method to analyze and discover attack scenario behind alerts. The assumption here states that the component attacks are usually not isolated, but related to different stages of the attacks, with the early ones preparing for the later ones. They introduce the notion of hyper alerts to represent the prerequisite and the consequence of each type of alert by using logical predicates. Each hyper-alert is a tuple (fact, prerequisite, consequence), where fact is the set of alerts attribute's names, and prerequisite and consequence are two different sets, each one consisting of a logical combination of predicates expressed as mathematical conditions on the variables contained in the set fact. The model employs distributed agents to collect alert information online and adopts prerequisite-consequence correlation method to analyse and discover attack scenario and intent intrusion behind the alerts.

An alert correlation method, based on causal approach, had also been proposed by (Zali at el, 2012). In this method, the knowledge base of attack patterns is represented as a graph model called causal relations graph. Some trees related to alerts probable correlations are constructed offline while the correlations of each received alert in real time with previously received alerts will be identified by performing a search only in the corresponding tree.

Alserhani at el (2010), developed a rule based correlation language MARS, a Multi-stage Attack Recognition System which is based on prerequisite-consequence-based correlation method to analyze and discover attack scenario behind alert. Unlike others, they add another two parameters for modeling attack consequences, i.e., vulnerability and extensional consequences. MARS is mainly based on the phenomena of cause and effect. It has two main components: online and offline. The main purpose of the online component is to receive raw alerts and generates hyper-alerts. Then, multi-stage attack recognition is applied to correlate hyper-alerts based on rules provided by the offline component

Ning et al. (2004), proposed alert correlation model based on prerequisites and consequences of individual detected alerts. A knowledge database "Hyper-alert Type Dictionary" contains rules that describe the conditions where prior behaviors prepare for later ones. Attack strategy is represented as a Directed Attack Graph (DAG) with constraints on the attack attributes considering the temporal order of the occurring alerts. The nodes of the DAG represent attacks and the edges represent causal and temporal relations. Similarities between these strategies are measured to reduce the redundancy. A technique of hypothesizing and reasoning about missing attacks by IDS is presented to predict attribute values of such attacks. The significance of their work is the reduction of the huge number of security incidents and to report a high-level view for the administrator. However, the proposed system is useful as a forensic tool where it perform offline analysis. In addition, building the knowledge database containing rules of the applied conditions is a burdensome wang at el (2008). However, authors have not provided a mechanism to build the Hyper Alert dictionary. Also, the generated graph is huge even with.

2.4 Expert System and Data Mining

Data mining is a process of discovering significant and potentially useful patterns especially in a large volume of data. Correlation mining is much effective because of the large number of correlation relationships among various kinds of suspicious alerts (karim at el, 2013). This method employs data mining algorithms on training data-set and using knowledge-base derived from human experts to identify attack scenarios on intrusion patterns and relationships among alerts. In this approach, statistical analysis of alerts can be done by identifying the co-occurrence of alerts within a predefined time window. Some relation rules or patterns will be created from correlation relationships that satisfy some statistical criteria. This involves pair-wise comparison between alerts since every two alerts might be similar and therefore can be correlated (Sadoddin at el, 2009). In this case, the repeated comparisons between alerts will lead to a very huge computational overload especially in largescale networks. Besides this, this approach requires a lengthy initial period of training (Mahboubian at el 2013).

A self-regulated alert correlation model had been proposed by (Yang at el, 2010). This model incorporated advantages of the associated component-based approach and alert information correlation based on preconditions. The model introduced data mining techniques in alert correlation and made improvements on alert correlation using improved Apriori algorithm.

In general, the Apriori algorithm states that any super-pattern of a non-frequent pattern is also not frequent (Zerin at el, 2011). In this proposed algorithm, it divided the alert information into several disjoint subsets with destination IP as a unit and then mined them separately before associating the correlation rules set for generating the results correlation rules set.

Alert correlation technique to automatically extract attack strategies from a large volume of intrusion alerts without specific prior knowledge about these alerts was proposed by (Zhu at el, 2007). The proposed approach is based on two different neural network approaches, which are multilayer perceptron (MLP) and support vector machine (SVM) to estimate the alert correlation probability by storing correlation strengths of any two types of alerts in an alert correlation matrix (ACM). For pattern recognition, SVM is a recommended classifier with its better performance (Phinyomark at el, 2011).

An algorithm to mine attack behaviour patterns from a large number of intrusion alerts without prior knowledge of

the attacks was proposed by (Kavousi at el, 2012). The proposed engine has two components. The first, an offline component that periodically learns multi-step attack behaviour patterns from historical alerts using a Bayesian causality analysis and the second is an online component that correlates alerts in real time using a hierarchical method and based on the attack behaviour patterns extracted by the offline component.

Zhitang at el (2008), employed different data mining algorithms for real-time correlation to discover multi-stage attacks. Off-line attack graph is constructed using manual or automatic knowledge acquisition and then attack scenarios are recognized by correlating the collected alerts in real-time. The incoming step of an attack can be predicted after detection of few steps of attack in progress. The association rule mining algorithm is used to generate the attack graph from different attack classes based on historical data. "Candidate attack sequences" are determined using a sliding window.

In Zhang, et al (2007), AprioriAll algorithm which is a sequential pattern matching technique is used to generate correlation rules based on temporal and content constraints. The work adopted a classical sequential mining method GSP to find the maximal alerts sequence and then to discover the attack strategy. The limitation of their work is the use of only attack class and temporal as features.

Cuppens et al. (2002), Proposed MIRADOR correlation approach for alert clustering, merging and then correlation. Explicit correlation of events based on security experts is used to express the logical or topologic links between events. Attack is specified using five fields and based on the language of LAMBDA. Partial matching techniques are adopted to build the model. In addition to explicit correlation, implicit correlation is used to overcome possibly missing events.

Qin at el (2005), proposed a combination of statistical and knowledgebase correlation techniques. Three algorithms are integrated based on assumption that some attack stages have statistical and temporal relations even though direct reasoning link is not existent. Bayesian-based correlation engine is used to identify the direct relations among alerts based on prior knowledge. In contrast to previous approaches, knowledge of attack steps incorporates as a constraint to probabilistic inference to avoid the exact matching of pre and post conditions. Causal Discovery Theory-based engine is developed to discover the statistical of one-way dependence among alerts. In addition, Granger Causality-based algorithm is used by applying statistical and temporal correlation, to identify mutual dependency. However, the problem of selection time window for temporal correlation is still an open problem. Attackers can exploit the slow-and low attack to avoid detection. Attack reduction also relies on prior knowledge where zero-day attack is not detected and also the analysis of results may result to huge computing work load Wang at el (2008).

3.0 DISCUSSION AND ANALYSIS OF ALERT CORRELATION TECHNIQUE

3.1 Comparison of Alert Correlation Techniques

All the discussed techniques have their advantages and disadvantageous as summarized in Table 1 below

Table 1: comparison of alert correlation techniques

Technique	Advantages	Disadvantages
Similarities of Alert Attributes technique	Can reduce large number of redundant alert generated by multiple sensors.	-Suitable for known alerts. -Not able to discover causality of alerts and statistical relationships. -Limited to discover complicated attacks.
Predefined Attack Scenario	is able to accurately detect well-documented attacks Can reduce large number of redundant alert generated by multiple sensors	Could generate large number of false positive alarm • it requires that users specify attack scenarios manually • It is limited to detection of known attacks or misuse detection and not anomaly detection. • multi-step attack alert is disregarded intrusion (Osman at el, 2010, Alsaidi, at el 2012, Benferhat at el, 2012)
Prerequisites and Consequences of Attacks	Multi-step attack can be detected to provide a high-level view of the attack associated with a security compromise • can generate useful graph to determine the attacker's objective	The approach may not be practical in production networks due to the complexity of the design and user behavior It is expensive to build a complete attack database which consists of every attack action with its pre- and post-conditions (Karunasekera at el, 2010, fatma at el, 2013).

data mining and Expert system	Does not need pre-defined knowledge about attack scenarios. • Using anomaly detection technique • new attack scenarios can be identified • can be used as pre-process alerts or meta-alert signatures.	a very huge computational overload especially in large scale networks This approach requires a lengthy initial period of training (Mahboubian at el 2013).
Hybrid technique	Performs multiple types of correlations (structure, cause & statistical) No predefined rules Recognize known and unseen alerts No manual parameters settings	May lead to complex architecture (Maheyzah at el 2015)

3.2 Proposed Design Consideration for Alert Correlation Technique

From the discussed alert correlation techniques, we have identified the significant issues within each technique which can be solved to improve the effectiveness of NIDS performance. Among the issues include alert Normalization and aggregation, alert correlation architecture, false alerts, alert prioritization, attack prediction, test data and the visualization techniques applied. This section briefly examines each issue and proposes a solution to fix it.

Table 2: Issues analysed in NIDS

Design consideration	Description	Proposed solution
Alert normalization	the majority of organizations implement different types of NIDSs (heterogeneous NIDSs), accordingly they produce alerts in different data format ((Maheyzah at el 2015, Rahayu at el, 2009)	Convert different alert data formats from multiple intrusion sensors into a standard format to be appropriate and acceptable by the other correlation components.
Attack scenario construction (study behavior of the attacker)	Attacks are likely to generate multiple related alerts. Current IDS do not make it easy for security officers to logically group related alerts (wang at el 2006; Xiao at el 2010)	To group or cluster alert which has a related event or event threaded. 2. classify the alerts into the corresponding cause effect paradigm
Alert correlation architecture (To solve problem of alert flooding)	IDS are prone to alert flooding as they provide a large number of alerts to the security officer, who then has the difficulties coping with the load (Bin at el, 2006,Rahayu at el, 2009)	To reduce number of alert generated from IDS and improve the alert correlation performance in terms of the processing time and quality of alerts by adopting alert filtration and alert aggregation
False Alert	Existing IDS are likely to generate false positives or false negatives alert (Rahayu at el, 2009)	To reduce number of false alerts through filtering 2. To identify known attack using misuse detection 3. To identify unknown attack using anomaly detection
Alert Severity/Prioritization	Not all generated alerts are equally important (Ghorbani at el2010; Maheyzah at el 2015, Alsubhi at el, 2008)	need to separate important alerts from the rest and calculate scoring and prioritizing alerts
Attack Prediction/ execution mode	Technologies are not effective in predicting the future attacks. (Maheyzah at el 2015; Wang at el, 2009; Rahayu at el 2008)	A proactive approach is to anticipate and conduct possible attacks to prevent damage
Test Data	The effectiveness of a component depends heavily on the nature of the data-set analysed to evaluate the system.	latest attack scenario data-sets to include IPv6 attack to ensure its efficiency and effectiveness in producing a good and quality output

Table 3: A summary of related works with design considerations

	Correlation application				Execution mode		Attack scenario		Correlation architecture		Test data set			
Alert correlation technique	Similarities	Predefined	Prerequisite	Expert system and data	Online	offline	Single	Multistage	Centralized	Distributed	selected	Darpa1999	Darpa2000	IPv4 or IPv6
Alserhania at el [2008]		✓	✓	✓		✓		✓	✓		✓		✓	
Bateni at el.[2013]		✓	✓	✓		✓		✓	✓		✓		✓	IPv4
Al-Saedi At el.[2012]	✓	✓	✓			✓	✓		✓			✓		IPv4
Mohamed at el.[2012]	✓				✓		✓			✓				IPv4
Osman at el[2012]	✓		✓			✓		✓	✓		✓		✓	
Anbarestani at el[2012].			✓	✓		✓		✓	✓				✓	
Zhaowen at el[2010]			✓		✓			✓		✓				IPV6
Zali at el[2012]			✓		✓			✓		✓			✓	
Kavousi at el [2012]				✓	✓			✓		✓			✓	IPv4
Mahboubian at el[2012]	✓	✓			✓			✓	✓				✓	IPv4
Amiri at el[2011].				✓		✓		✓		✓				
Ebrahimi at el [2011]		✓				✓		✓	✓				✓	
Marchetti at el [2011]		✓		✓		✓		✓	✓					IPv4
Taha at el[2010]		✓		✓		✓		✓		✓	✓	✓	✓	IPv4
Tabia at el [2010]		✓		✓		✓		✓	✓					IPv4
Yang at el [2010]		✓	✓			✓		✓	✓		✓			IPv4
Tian at el [2009]	✓		✓			✓		✓		✓	✓		✓	
Huang at el[2010]		✓		✓				✓	✓				✓	IPv4
Ahmadinejad at el [2009]	✓					✓		✓	✓				✓	IPv4
Valdes at el [2008]	✓					✓		✓	✓				✓	
Ning at el [2004]	✓					✓		✓	✓				✓	
Zhitang at el [2008]	✓					✓		✓	✓				✓	
Zhang	✓					✓		✓	✓				✓	
Wang at el 92008)	✓					✓		✓	✓				✓	

An efficient correlation system should achieve the following objectives

i. Alert normalization to convert different alert data formats into a standard format
ii. Reducing and eliminating redundant of intrusion alerts
iii. Ability to discover complete attacker strategy with known and unknown attack
iv. A unified hybrid architecture that leverages capabilities of the various correlation techniques
v. Filtering and prioritizing intrusion alerts to improve the quality of alerts.
vi. A proactive approach to predict the next attacker action in real time
vii. graphical based approach for analysis and presenting alerts
viii. Take different types of dataset to measure components effectiveness

4. CONCLUSION

Several techniques of alert correlation have been proposed to help identify and discover the relationships amongst alerts from multiple sources. However, most of these techniques suffers from complex correlation rules definition, they have limited capabilities to discover new and complicated attack, depends on human expert's knowledge to update the correlation knowledge as well as they do not provide a proactive action when attack activities are going on. As a result researchers have begun to look for a hybrid approach that leverages capabilities of the various correlation techniques

Future work should be based on scalable, structured and computationally techniques which do not require prior knowledge, not dependable on security expert to frequently update rules and are able to detect known and unknown attacks. Additionally, as the Internet enters a new era and domain such as mobility and Internet of Things and its ever-growing user's base, a more flexible and intelligent intrusion alert which is able to detect and predict the incoming alerts at sensor level and in real time is desired to complement IDSs to secure information systems, networks and sensitive data

5. REFERENCES

1. Alserhani, F., Akhlaq M., Awan I.U., Cullen A.J., Mirchandani P., (2010). "MARS: Multi-stage Attack Recognition System", 24th IEEE International Conference on Advanced Information Networking and Applications (AINA)
2. Zhaowen Lin, Shan Li and Yan Ma,(2010). "Real-Time Intrusion Alert Correlation System Based on Prerequisites and Consequence", 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM)
3. Zhi-tang Li, Jie Lei, Li Wang, Dong Li, 2007,"A Data Mining Approach to Generating Network Attack Graph for Intrusion Prediction," Fuzzy Systems and Knowledge Discovery, Fourth International Conference on, vol. 4, pp. 307-311, Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007) Vol.4
4. Ai-fang Zhang, Zhi-tang Li, Dong Li, Li Wang, 2007"Discovering Novel Multistage Attack Patterns in Alert Streams," Networking, Architecture, and Storage International Conference on Networking, Architecture, and Storage (NAS 2007)
5. Jie Ma, Zhi-tang Li, Wei-ming Li, 2008."Real-Time Alert Stream Clustering and Correlation for Discovering Attack Strategies," Fuzzy Systems and Knowledge Discovery, Fourth International Conference on, vol. 4, pp. 379-384, 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery,
6. PengNing, Yun Cui, Douglas S. Reeves, 2002"Constructing Attack Scenarios through Correlation of Intrusion Alerts," in Proceedings of the 9th ACM Conference on Computer & Communications Security, pages 245--254, Washington D.C.
7. P. Ning, D. Xu, C. Healey, R. St. Amant,2004 "Building Attack Scenarios through Integration of Complementary Alert Correlation Methods," in Proceedings of the 11th Annual Network and Distributed System Security Symposium
8. X. Qin. 2005A Probabilistic-Based Framework for INFOSEC Alert Correlation. PhD thesis, Georgia Institute of Technology
9. X. Qin and W. Lee2004. Attack plan recognition and prediction using causal networks. In ACSAC '04: Proceedings of the 20th Annual Computer Security Applications Conference (ACSAC'04), pages 370-379, Washington, DC, USA,. IEEE Computer Society.
10. W.Wang, T.E.Daniels,(2008). A graph based approach toward network forensics analysis, ACM Transactions on Information and Systems Security
11. Li Wang, Ali Ghorbani, and Yao Li Automatic Multi-step Attack Pattern Discovering,International Journal of Network Security, Vol.10, No.2, PP.142{152, Mar. 2010
12. Siti Rahayu Selamat, R. S. (2008). Mapping Process of Digital Forensic Investigation Model. *IJCSNS International Journal of Computer Science and Network Security* , Vol. 8(No. 10): p. 163-169.
13. Alharbi, S. e. (2011). The Proactive and Reactive Digital Forensics Investigation ProcessInternational Journal of Security and Its Applications Vol. 5 No. 4, October, 2011
14. pandaLabs, Annual Report Panda Security's AntiMalwa re Laboratory 2009. 2010, Panda Security
15. Siti Rahayu Selamat, Shahrin SahibA Forensic Traceability Index in Digital Forensic Investigation ,Journal of Information Security, 2013, 4, 19-32
16. Ricci S.C, I. F. (2006). Digital forensics investigation model that incorporate legal issues. *Digital Investigation* , 29-36
17. S. Garfinkel, "Anti-forensics: Techniques, detection and countermeasures," in 2nd International Conference on i-Warfare and Security, 2007, p. 77.
18. Sebastian Roschke, Feng Cheng, and ChristophMeinel (2011), A New Alert Correlation Algorithm Based on Attack Graph, In Proceedings of the 4th International

Conference on Computational Intelligence in Security for Information Systems (CISIS 2011), Torremolinos,
19. Wang, L., Liu, A., Jajodia, S (2006): Using attack graphs for correlation, hypothesizing, and predicting intrusion alerts. Journal of Computer Communications
20. FatmahBahareth, OmaimaBamasak, 2013, Constructing Attack Scenario using Sequential Pattern Mining with Correlated Candidate Sequences. The Research Bulletin of Jordan ACM,
21. C. V. Zhou, C. Leckie, and S. Karunasekera. 2010,"A survey of coordinated attacks and collaborative intrusion detection,"Comput. Secur., Vol. 29, no. 1, pp. 12440,
22. RobiahYusof, ,SitiRahayuSelamat, Shahrin Sahib 2008, Intrusion Alert Correlation Technique Analysis for Heterogeneous Log, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.9K. H. Al-Saedi, S. Ramadass, A. Almomani, S. Manickam, and W. A. A. Alsalihy, 2012, collection mechanism and reduction of IDS alert, International journal for Compututer Application., Vol. 58, no. 4
23. A. B. Mohamed, N. B. Idris, and B. Shanmugum, 2012, Alert correlation using a novel clustering approach, in International Conference on Communication Systems and Network Technologies (CSNT), Gujarat, India, May 1113, pp. 7205.
24. H TagelsirElshoush, Izzeldin Mohamed Osman, 2012,An improved framework for intrusion alert correlation, in The World Congress on Engineering, London, UK, Jul. 46, pp. 51823
25. TagelsirElshoush, Izzeldin Mohamed Osman, 2011, Alert correlation in collaborative intelligent intrusion detection systems—a survey, *Applied Soft Computing*, Vol. 11, 7, pp. 434965
26. A. Ebrahimi, A. Z. H. Navin, M. K. Mirnia, H. Bahrbegi, and A. A. A. Ahrabi, 2011, Automatic attack scenario discovering based on a new alert correlation method, in IEEE International Systems Conference M. Marchetti, M. Colajanni, and F. Manganiello, 2011, Identification of correlated network intrusion alerts, in Third International Workshop on Cyberspace Safety and Security (CSS), Milan, Italy, pp. 1520.
27. P. Ning, Y. Cui, D. S. Reeves, and D. Xu,2004, .Techniques and tools for analyzing intrusion alerts, ACM Trans. Inf. Syst. Secur. (TISSEC), Vol. 7, no. 2, pp. 274318
28. Z. Zali, M. R. Hashemi, and H. Saidi, 2012,"Real-time attack scenario detection via intrusion detection alert correlation,"in 9th International ISC Conference on Information Security and Cryptology (ISCISC), Tarbiz, Iran, Sep. 1314, , pp. 95102
29. M. Karim, C. F. Ahmed, B. S. Jeong, and H. J. Choi, 2013, "An efficient distributed programming model for mining useful patterns in big datasets," IETE Tech. Rev., Vol. 30, no. 1, pp. 5363
30. Reza. Sadoddin, and A. A. Ghorbani,"An incremental frequent structure mining framework for real-time alert correlation," Comput. Secur., Vol. 28, no. 3, pp. 15373, May 2009
31. M. Mahboubian, N. I. Udzir, S. Subramaniam, and N. A. W. A. Hamid, 2012An alert fusion model inspired by artificial immune system, in International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec), Kuala Lumpur, Malaysia
32. L. Yang, and D. Xinfa, 2010.Alert correlation model design based on self-regulate, in Second International Conference on Multimedia and Information Technology (MMIT), Kaifeng, China,
33. S. F. Zerin, and B. S. Jeong 2011, A fast contiguous sequential pattern mining technique in DNA data sequences using position information, IETE Tech.
34. Bin Zhu and Ali A. Ghorbani, 2006 Alert Correlation for Extracting Attack Strategies, International Journal of Network Security, Vol.3, No.3, PP.244–258,
35. A. Phinyomark, P. Phukpattaranont, and C. Limsakul2011,"A review of control methods for electric power wheelchairs based on electromyography signals with special emphasis on pattern recognition,"IETE Tech. Rev., Vol. 28, no. 4
36. F. Kavousi, and B. Akbari, 2012Automatic learning of attack behavior patterns using Bayesian networks, in Sixth International Symposium on Telecommunications (IST), Tehran, Iran
37. C. J. Huang, C. F. Lin, C. Y. Li, J. J. Liao, Y. W. Wang, and K. W. Hu, 2010An adaptive rule-based intrusion alert correlation detection method, in First International Conference on Networking and Distributed Computing (ICNDC), Hangzhou, China
38. B. Abdulrazak, and Y. Malik, 2013Review of challenges, requirements, and approaches of pervasive computing system evaluation, IETE Tech .Rev., Vol. 29
39. ShameliSendi A, Dagenais M, Jabbarifar M, Couture M. 2012Real time intrusion prediction based on optimized alerts with Hidden Markov model. Journal of Networks.
40. Alsubhi K, Al-Shaer E, Boutaba R., 2008, Alert prioritization in intrusion detection systems. Proceedings of IEEE Network Operations and Management Symposium. NOMS.
41. Ghorbani AA, Lu W, Tavallaee M.,2010, Alert management and correlation. Journal of Network Intrusion Detection and Prevention..
42. Ning P, Cui Y, Reeves DS, Xu D. 2004Techniques and tools for analyzing intrusion alerts. ACM Transactions on Information and System Security (TISSEC)..
43. Bateni M, Baraani A, Ghorbani, 2012,A. Alert correlation using artificial immune recognition system. International Journal of Bio-Inspired Computation..
44. Xiao F, Jin S, Li X. 2010, A novel data mining-based method for alert reduction and analysis. Journal of Networks..
45. MaheyzahMdSiraj, Hashim Hussein TahaAlbasheer and Mazura Mat Din,2015,Towards Predictive Real-time Multi-sensors Intrusion Alert Correlation Framework Indian Journal of Science and Technology, Vol 8(12)

International Journal of Computer Applications Technology and Research
Volume 5– Issue 4, 234 - 237, 2016, ISSN:- 2319–8656

New Proposed Mobile Telecommunication Customer Call Center Roster Scheduling Under the Graph Coloring Approach

Hasitha Indika Arumawadu
School of Computer Science and Technology, Wuhan University of Technology, Wuhan, P. R. China

R. M. Kapila Tharanga Rathnayaka
Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuoya, Sri Lanka

D. M. K. N. Seneviratna
Faculty of Engineering, University of Ruhuna, Galle, Sri Lanka

Abstract: The call center roster scheduling is one of the significant problem in the mobile telecommunication roster management systems today; especially, creates work plan and allocates working hours for the whole day under the three shifts creates big challenge for the administrators who responsible for creating roster time tables. As a result of assigning employees into roster timetables under the manual scheduling systems create this problem more complicated. This new proposed automated roster scheduling approach developed under the two stages. As an initially, Enhanced Greedy Optimization algorithm is implemented to optimize the hotline roster and compared with other optimization algorithms (Simulated Annealing and Genetic Algorithm). In the Second stage, client server based framework introduced to access and update roster timetables for administrators as well as employees with different access levels.
Keywords: Graph Coloring; Genetic Algorithm; Hotline Roster Scheduling; Mobile Telecommunication; Call Center

1. INTRODUCTION

A call center is a centralized office used for receiving or transmitting a large volume of requests by telephone. It is a central point from which all customer contacts are managed and provides assistance for the customers to solve their problems related to the service offered by the company. It is generally a part of company's customer relationship management. It can be considered as a big challenge in the modern world today; especially in telecommunication industry under the 24 hour service over the 365 days. Currently, even in most of the reputed companies, spend additional time for creating working schedule manually [1].

The feasible working schedule has mainly impact on the quality of service for their customers as well as their employers. As a result, make unbiased manual scheduling timetable in telecommunication industry has been created more complicated problems today. Moreover, various necessities of customers, different type of qualifications, experiences and specializations, employers and employee requirements, unpredictable incidence and absenteeism and other factors make the problem more complicated.

Before make feasible Customer care hotline roster management system, we should have to concern much more about constraints and requirements going under the different shifts as morning, afternoon and night [2]. Based on customer's feedbacks, the afternoon session has identified as peak and Night and Morning considered as off-peak time. So, many customers take assistance over the phone in Peak time comparing to off-peak time. As a multi ethnic country, especially in Sri Lankan top level companies has been offering different languages in their hotline systems. So, language skills of their customer officers are another requirement we should consider.

Generally, customers can be divided into two categories as loyalty customers and normal customers. To assist loyalty customers, need to assign more experienced and skilled officers to spread balance way to serve better service. Furthermore, some unpredictable incidents and absenteeism should be also consider; especially different religious officers have different holidays, so have to consider religious wise holidays before assigning them into rosters.

In this scenario manual system for scheduling timetable is not effective, flexible and balanced with advanced than in IT technology and efficient optimization algorithms [3]. So, our proposed work, focus on developing a framework based on Graph Coloring algorithm for customer care hotline management system especially for highly populated companies in telecommunication industry in Sri Lanka.

In our approach, client server architecture is proposed for retrieve and manage the employee information. The rest of the paper is organized as follows. Section 2 explains about brief overview of existing solutions with pros and cons. Section 3 explains about proposed work with enhanced greedy optimization algorithm. Section 4 explains about experimental results and Section 5 ends up with conclusion and future work.

2. LITERATURE REVIEW

In the last two decades, Graph coloring theories have been widely applied in the literature to find the solutions for real world problems; especially, timetable scheduling, nurse roster scheduling, register allocation, problem of bandwidth allocation and etc [4]. Timetabling is the process of assigning limited resources to a set of events without violating the constraints[5][6].Most of the current proposed solutions either make use of random based optimization algorithms which won't be efficient or applicable only for fully automated scheduling problems.

Kundu & Mahato (2007) have done remarkable studies based on Genetic Algorithm (GA) and had deep discussions about the role of data warehouse management system to handle hospital and nurse management information. Based on soft and hard constraints and simulated Annealing, they have described the use of Genetic Algorithm (GA) for solving NSP under the three different levels [7]. Bard & Purnomo (2005) reported a same kind of methodology to solve the nurse roster problem using column generation. This sub problem was formulated as a shortest path problem with resource constraints, where each possible shift was represented by a node and it was solved by using a two stage algorithm [8].

Hussain et al.(2011) reported a method to solve the exam timetabling problem based on column generation with graph coloring approach and successfully applied for clustering heuristic to determine the solution quality for exam timetabling problem[9]. In the same time, Razak (2010) has used bipartite graph edge coloring approach to course timetabling. The problem was modeled based on two graphs. The vertices in the first graph represent the lectures who teach a class and the vertices in the second graph are groups of students attending a class and edges represent hard constraint or the meet up between the groups of students attending the class and the lectures who teach the class's respectively [3].

Elghazelet al. (2006) have proposed a new approach of clustering based on a b-coloring of graphs to define a typology of patients. Their contribution in this work was a new clustering algorithm. It was used pair-wise representation, where the objects (hospital stays) are mapped to the nodes of an undirected edge-weighted graph, where the edge weights reflect the dissimilarity between the corresponding pair of nodes. The clustering problem was then formulated as a graph coloring problem [10].

However, Mouna et al. (2011) introduced a new concept for developing the multi-level graph coloring approach and successfully applied it for Bus Driver's timetables. In this context, first objective is to minimize the break time between two travels of one driver. The second objective was working informing between all services. Finally the third objective consists in minimizing the total working hours of all services [11].

3. PROBLEM DEFINITION

Huge companies such as telecommunication companies have continuous service all over the day. Comparing with human work capacity, it should divide into three shifts. In such scenario, assign work schedule to each worker and create timetable is challenging task. Most of the current research works has been proposed with either random based optimization algorithms or local optimization algorithms which cannot be used in such difficult scenarios. Especially there are some unpredictable situations such as workers may change their current shifts due to emergencies. In addition uninformed absenteeism, different demand of customers, different religious workers, balancing experience level and skill level of the workers makes the problem more complicate[12][13]. In such kind of scenarios the one who responsible for making timetable has to face difficulties. In our proposed work we can dynamically manage these kinds of especial scenarios.

Our main target is to find high quality unbiased feasible scheduling algorithm and resource assignments under the labor contract rules and satisfying employees as well as employers' requirements and constraints. Here we list down constraints that we were considered.

Constraints

- Each Customer Officer required working a shift per day.
- Each Customer Officer gets at least one day off per week.
- Each Customer Officer entitled to one day off after a night shift.
- Minimum number of Customer Officers in Peak Time is 8 and Off Peak Time is 5.
- Minimum number of Senior Customer Officers in Peak Time is 3 and Off Peak Time is 2.
- At least one Head Customer Officer in a shift.
- Each Head Customer Officers entitled at least two day off per week.
- Assign skilled Customer Officers to handle Loyalty customers.
- Assign different language Customer Officers per each shift.
- Customer Officers may go on holiday and will not work shift during this time.
- Different religious Customer Officers have different holidays.
- Balancing the work load among Customer Officers.
- Increase maximum number of customer Officers in special holiday days like New Year day, Christmas day etc.

In our proposed work C#.net as a programming language and SQLServer for handle database management system were used.

4. METHODOLOGY

Customer Care roster scheduling represents the important administrative activity in top level telecommunication companies today. Major task is to identify the main areas, main working categories and allocation of recourses in efficient way. Based on requirement analysis, five main categories were identified. They were administrators (employer), head customer officer and other customer officers (employee), customers and other staff members.

4.1 Client Server Architecture: Roster Analysis Framework

The new proposed system mainly based on client server architecture. The Employers information stored in database and server is responsible for handles the request of the client to complete some logic task. So, administrator and others uses can directly login to the system through the interface provided and can update and view their records through the system.

As a first step, Administrator in the system is a main user. Administrator is a main handler who controls this system. For reliable handing, administrator appointed sectional heads or Head Customer Officers for each and every department. Sectional heads have a responsibility to maintain their departments' accounts. As a main user, Head Customer Officers creates new accounts for other customer officers in his department. Hence, all the others must register the system

and need to create their own accounts. Once user gets activated their access, system provides a facility to update their details, apply leaves, apply and confirm OT, view their time tables. So, members must send their leave details through their accounts before creating the roster. In this proposed system, Modified Greedy algorithm is used to create feasible high quality roster.

4.2 Enhanced Greedy Optimization Algorithm

In general, a greedy algorithm that follows the heuristic problem solving with locally optimal choice at each stage with the hope of finding a global optimum. Furthermore, it may yield locally optimal solutions that approximate a global optimal solution in a reasonable time. The new Proposed Enhanced Greedy Algorithm has followed five crucial components. They are; Candidate set should be created, selection function should be selects the best candidate to add to the solution, feasibility function should be defined to identify whenever a candidate can be used to contribute to a solution, modified objective function is used to assign a value to local solution, or a partial solution and Solution function helps to identify the complete solution of the intended problem.

The new Proposed Greedy algorithm can be divided into two categories namely; (1) graph coloring approach based ordering for scheduling (2) Clustering [14, 15, 16 and 17].

From a mathematical point of view a graph does not need to be drawn. If we have large number of vertices and edges, it may be make big problem for draw graph. So in real life problems, mathematical methods were used.

One of the definition of graph G, G= (V (G), E (G)) consist of two finite sets. V (G), the vertex set of the graph G, often denoted by just V, which is a non-empty set of elements called vertices. E (G), the edge set of the graph, often denoted by just E, which a possibly empty set of elements is called edges.

Graphical coloring can be simply explained as coloring the nodes of a graph with the minimum number of colors. In graph theory two adjacent vertices cannot be colored using the same color. The term chromatic number $\chi(G)$ is defined as the minimum number of colors needed for coloring of G. A graph G is k chromatic, if $\chi(G) = k$ and *G* is *k* colorable, if $\chi(G) \leq k$.

Below algorithm used for create adjacency Matrix according to the constraints.

```
create array customer_officer
get array length array_length
create matrix

for (i=0 to array_length) {
  for (j=0 to array_length) {
     if ( customer_officer[i]==customer_officer[j] ) {
              matrix[i][j]=1;
     }
      else {
               matrix[i][j]=0;
      }
    }
     }
```

Below algorithm used for create the graph using above adjacency matrix.

```
{
Vertex (G) {x1,x2…….xn}
Colour(G) {c1, c2,c3,c4…..cn}
for (i=0; i<n; i++) // number of vertices
{
for each colour vertex u -> n2(vi) do
{
TabooColors(color(u)) = vi
}
Colour (vi) = min{c :TabooColors(c) !=vi}
loop: Let Ci be the first colour in C.
For each j with i<j and xj adjacent to xj in G
Set Cj = Cj-{ Ci} //where xjwill not have same colour as xi
Change i to i+1 and if i+1 ≤ n,
Return to “loop”.
}
```

Optimized solution of the scheduling can be generated with set of feasible solutions and optimization function. At each generation, decision should be best with respect to time.

5. EXPERIMENTAL RESULTS AND DISCUSSION

Table 1. Comparison results for selected optimization algorithms

Days	Count	Algorithm	Solved	Avg. Time (Sec.)
7	1000	PA	920	1.34
		SA	880	1.54
		GA	800	5.00
14	1000	PA	940	2.68
		SA	920	5.70
		GA	730	10.2
21	1000	PA	940	6.41
		SA	910	6.96
		GA	860	15.4
30	1000	PA	970	21.2
		SA	650	23.5
		GA	640	24.1

[a.] Simulated results from different Algorithms

Figure 1. Admin console and grouping interface

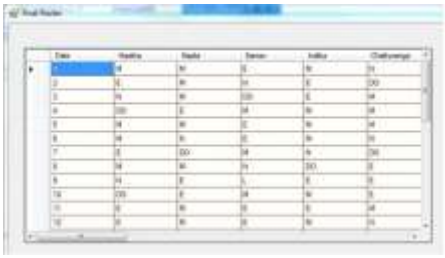


Figure 2. Final roster output

C#.net programming is used to access information from database. As an initial step, proposed Greedy optimization algorithm is implemented in C#.net. Detailed information system is created and accessed through SQLServer.

Moreover, optimization with predefined cost function and set of feasible solutions are used to determine the best optimized value. The Problem size is considered as 1000 and compared the proposed greedy solutions with simulated annealing and Genetic algorithm. Optimization is done in weekly in four terms of 7, 14, 21 and 30 days.

Table 1 presents the comparison of proposed results (PA) with simulated annealing (SA) and genetic algorithm (GA) with respect to time in seconds.

As an initial step, average time taken for our proposed greedy is 1.34 sec whereas for SA it is 1.54 and GA it is 5.00 seconds. Next, 14 days PA has taken 2.68 sec whereas SA has taken 5.70 and 10.2 for GA respectively. Similarly, for 21 day schedule, 3.21 is for PA, 5.70 for SA and 8.26 for GA. Compared with first two weeks simulation results, third and fourth week simulations result is almost same for PA, SA and GA. According to the results, all the three optimization algorithms holds good to get optimized schedule results in long period simulations. So we can conclude that, it is better to choose greedy or simulated annealing for short period scheduling and can use genetic optimization for long period scheduling's.

6. CONCLUSION

Customer hotline roster scheduling is the one of the crucial problem in current telecommunication field. Different kind of special constrains make the problem more and more complicated. Here we proposed new client server roster scheduling framework using graph coloring approach. Proposed frame work is easy to maintain and easy to update for administrator and employee.

In future, enhancements like linking all the databases in different branches for manage all the rosters in every branch in one place can be done to make it easy and high quality.

7. REFERENCES

[1] A. Hon, W. Chun, S. Ho, C. Chan, G. Pui, S. Lam, F. Ming, F. Tsang, and J. Wong, Nurse Rostering at the Hospital Authority of Hong Kong, 2000.

[2] S. K. Amponsah, E. Agyeman, and K. G. Okrah, Graph Colouring , an Approach to Nurses Scheduling , Case Study : Ejura District Hospital , Ashanti Region , Ghana, vol. 6, no. 1, pp. 1–5, 2011.

[3] H. A. Razak and Z. Ibrahim, Bipartite Graph Edge Coloring Approach to Course Timetabling, pp. 229–234, 2010.

[4] A. E. P. D, Graph-Coloring for Course Scheduling – A Comparative Analysis based on Course Selection order, pp. 83–88, 2014.

[5] G. L. Prajapati, A. Mittal, R. I. D. Yhuwh, F. Rswlpl, and D. Frpsohwhqhvv, An Efficient Colouring of Graphs Using Less Number of Colours, pp. 666–669, 2012.

[6] A. Elhag and E. Özcan, Expert Systems with Applications A grouping hyper-heuristic framework : Application on graph colouring, Expert Syst. Appl., vol. 42, no. 13, pp. 5491–5507, 2015.

[7] S. Kundu, M. Mahato, B. Mahanty, and S. Acharyya, Comparative Performance of Simulated Annealing and Genetic Algorithm in Solving Nurse Scheduling Problem, vol. I, pp. 19–21, 2008.

[8] J. F. Bard and H. W. Purnomo, Preference scheduling for nurses using column generation q, vol. 164, no. 2, pp. 510–534, 2005.

[9] B. Hussin, A. Samad, and H. Basari, Exam Timetabling Using Graph Colouring Approach, pp. 139–144, 2011.

[10] H. Elghazel, H. Kheddouci, V. Deslandres, A. Dussauchoy, and C. Bernard, A New Graph-Based Clustering Approach : Application to PMSI Data.

[11] M. Mouna, Y. MASMOUDI, and H. C. GIAD, A Multi-level Graph Coloring Approach for the Bus Driver ' s Timetables :, pp. 25–31, 2011.

[12] R. M. K. T. Ratnayaka, Z. Wang, S. Anamalamudi, and S. Cheng, Enhanced Greedy Optimization Algorithm with Data Warehousing for Automated Nurse Scheduling System, vol. 2012, no. December, pp. 43–48, 2012.

[13] I. Journal and C. Science, AUTOMATED SYSTEM FOR NURSE SHEDULING USING GRAPH, vol. 2, no. 3, pp. 476–485, 2011.

[14] Rathnayaka, R.M.K.T. ; Wei Jianguo ; Seneviratna, D.M.K.N., Behavior, Economic and Social Computing (BESC), 2014 International Conference on DOI: 10.1109/BESC.2014.7059517

[15] RM Kapila Tharanga Rathnayaka, DMKN Seneviratna, Wei Jianguo, "Grey system based novel approach for stock market forecasting", Grey Systems: Theory and Application, 5(2), 178-193, 2015, Emerald Group Publishing Limited.

[16] Hasitha Indika Arumawadu, RM Kapila Tharanga Rathnayaka, SK Illangarathne, "K-Means Clustering For Segment Web Search Results", International Journal of Engineering Works, 2(8), 79-83, 2015, kwpublisher.com.

[17] Hasitha Indika Arumawadu, RM Kapila Tharanga Rathnayaka, SK Illangarathne, "Mining Profitability of Telecommunication Customers Using K-Means Clustering", Journal of Data Analysis and Information Processing, 3(3), 63, 2015.