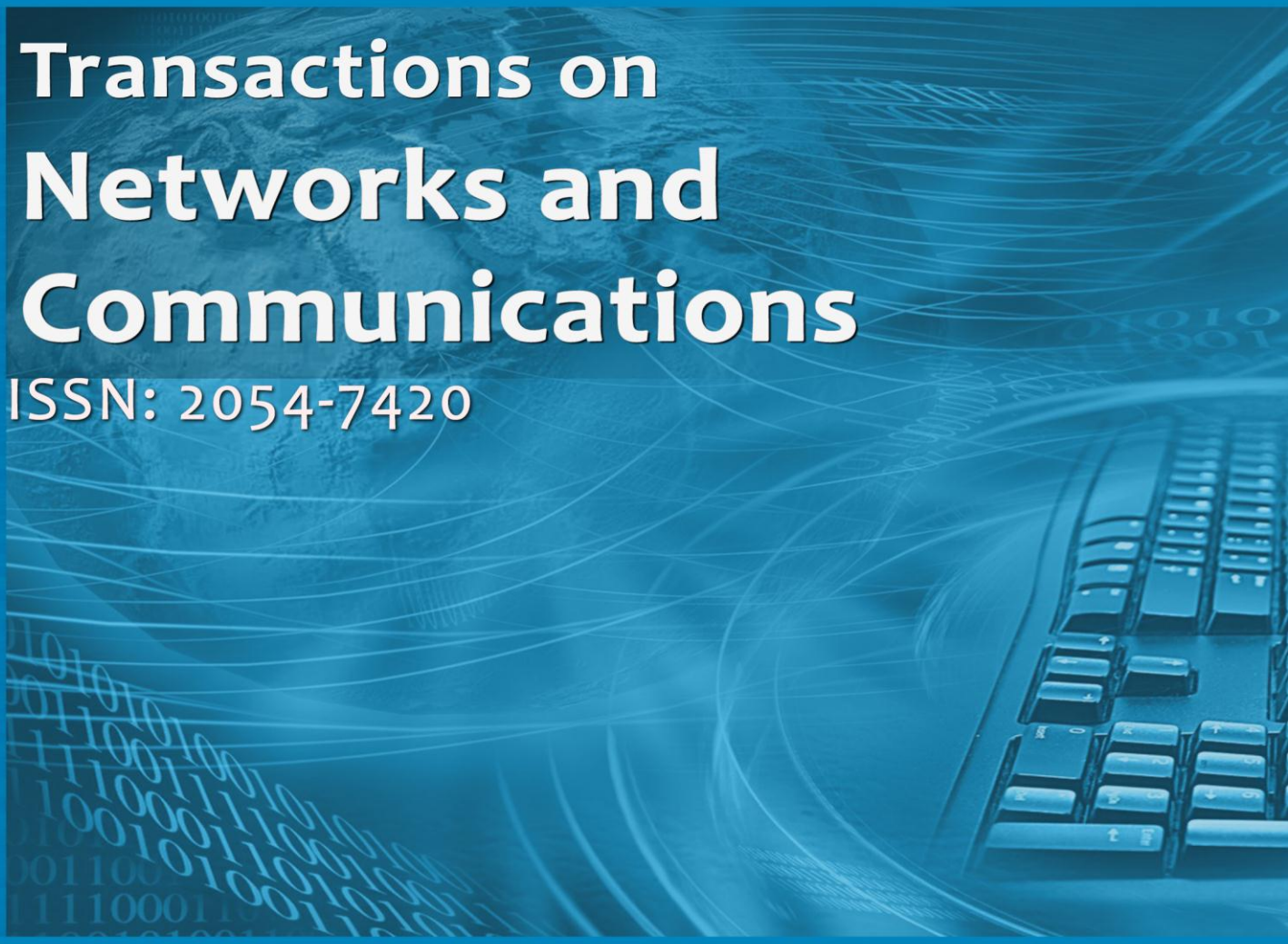
Volume 3. Issue 2 - April 2015

Transactions on Networks and Communications

ISSN: 2054-7420



Society for Sciences and Education - United Kingdom

TABLE OF CONTENTS

EDITORIAL ADVISORY BOARD

DISCLAIMER

All the contributions are published in good faith and intentions to promote and encourage research activities around the globe. The contributions are property of their respective authors/owners and the journal is not responsible for any content that hurts someone's views or feelings etc.

TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

Breast Cancer Risk Prediction Using Data Mining Classification Techniques

[1]Peter Adebayo Idowu, [2]Kehinde Oladipo Williams, [3]Jeremiah Ademola Balogun and [4]Adeniran Ishola Oluwaranti

[1, 3, 4]Department of Computer Science and Engineering, Faculty of Technology, Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria

[2]Department of Physical and Computer Sciences, College of Natural Applied Sciences, McPHERSON University, Ajebo, Ogun State, Nigeria

kehindewilliams@yahoo.com; paidowu1@yahoo.com

ABSTRACT

Breast cancer poses serious threat to the lives of people and it is the second leading cause of death in women today and the most common cancer in women in developing countries in Nigeria where there are no services in place to aid the early detection of breast cancer in Nigerian women. A number of studies have been undertaken in order to understand the prediction of breast cancer risks using data mining techniques. Hence, this study is focused at using two data mining techniques to predict breast cancer risks in Nigerian patients using the naïve bayes' and the J48 decision trees algorithms. The performance of both classification techniques was evaluated in order to determine the most efficient and effective model. The J48 decision trees showed a higher accuracy with lower error rates compared to that of the naïve bayes' method while the evaluation criteria proved the J48 decision trees to be a more effective and efficient classification techniques for the prediction of breast cancer risks among patients of the study location.

Keywords: breast cancer, classification, prediction, risk factors, naïve bayes, J48 decision trees

1 Introduction

According to WHO (2002) cancer has been responsible for the deaths of millions of people worldwide with an estimated increase of 50% for developing countries and for 70% of the total deaths due to cancer. According to Parkin et al (2003) developing nations only possess 5% of global funds for cancer control and very few human and material resources are also available in such countries (Grey et al, 2006).

The American Cancer Society (2008) defines cancer as a generic term for a large group of diseases that can affect any part of the body; other terms are malignant tumors and neoplasm. Breast cancer is a type of cancer which affects the breast tissue which is most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk (Sariego, 2010). Breast cancer is caused by a number of factors called risk factors; they are classified as either modifiable (those that can be controlled like habits, environmental hazards, etc) or non-modifiable factors (those that cannot be controlled like, gender, family history etc). According to the Collaborative Group on Hormonal Factors in Breast Cancer (2002), the primary risk factors for breast cancer are being female and of an older age. Other potential risk factors include: family history of breast cancer, age of menarche (first occurrence of menstruation), age of first birth, age of menopause, body weight (BMI), alcohol

DOI: 10.14738/tnc.32.662
Publication Date: 10th March 2015
URL: http://dx.doi.org/10.14738/tnc.32.662

consumption, exposure to radiation (Poongodi et al, 2011), higher hormonal levels and diet (Yager, 2006).

According to Johnson et al (2009) smoking tobacco appears to increase the risk of breast cancer which is higher depending on how long the person has been smoking. Long term smokers have an increased risk of about 35% to 50% (Santoro, 2009). The risk of breast cancer increases with an increased diet especially for those with fat diet (Blackburn, 2007), alcohol intake (Bofetta et al, 2006) and obesity. Radiation exposure (American Cancer Society, 2005) also increases the chances of breast cancer risk especially for women who have yearly mammogram tests especially between the ages 40 to 80 years face a risk of 225 in every million women screened (Hendrick, 2010). Also, exposure to pesticides, chemicals and organic solvents are believed to increase breast cancer risks also (Ferro, 2012). According to Boris et al (2010) genetics is believed to be the cause of 5% to 10% of breast cancer cases with those with none, one or two affected relatives with breast cancer before the age of 80 has a 2.3%, 4.2% and 7.6% risk respectively (Gage et al, 2012). Those with first degree relative with the disease face double the risk than a normal person.

Breast cancer risks can be reduced via early detection of the disease; according to the American Cancer Society (2007) early detection of breast cancer risks can help reduce the possibility of mitigating the full growth of tumors. The various ways of detecting breast cancer may include: clinical examination by a physician, self breast examination and mammography. Clinical examination of breast by a physician is one of the effective ways of reducing breast cancer mortality; it is required that a woman goes for clinical examination annually when above 40 years and every 3 years when between 20 and 40 years. Mammography involves the use of x-rays but with lower radiation; it has a breast cancer detection accuracy of 85 to 90% where routing mammogram leads to a 25 to 30% decrease in breast cancer mortality (American Cancer Society, 2007). Self-breast examination involves monthly observation of the breast and underarm by the patient; it allows the patient to be familiar with her breast and easily detect any anomaly she observes during the exercise. Diagnosis is the process of predicting the presence of breast cancer as either benign or malignant cases.

Classification is a data mining technique which involves the use of supervised machine learning techniques which assigns labels or classes to different objects and groups. It involves the process of model construction (analysis of training data for patterns) and model usage where the constructed model is used for classification. Classification accuracy is usually estimated as the percentage of test samples that are correctly classified.

This study aims at using data mining techniques to classify breast cancer risks using datasets of patients' information from LASUTH which contains the risk factors and the cancer classes (unlikely, likely and benign). The J48 decision trees and naïve bayes' classification of breast cancer was performed using the WEKA software.

2 Related Works

A number of papers have been documented and published on the use of data mining techniques in the classification of breast cancer risks. Some of such works are reviewed in the following paragraphs.

According to Rajesh et al (2012) who used SEER dataset for the diagnosis of breast cancer using the C4.5 classification algorithm. The algorithm was used to classify patients into either pre-cancer stage or potential breast cancer cases. Random tests were performed on the dataset which contained information for 1183 patients including the age of diagnosis, regional lymph nodes measures, and sequence number of tumors, dimension of primary tumor and contiguous growth of the primary

tumor. The analysis involved the use of three random 500 records form the pre-processed data of 1183 and was used as training data and the lowest error rate achieved was 0.599. During the testing phase, the C4.5 classification rules were applied to a test sample and the algorithm showed had an accuracy of 92.2%, sensitivity of 46.66% and a specificity of 97.4%. Future enhancement of the work will require the improvisation of the C4.5 algorithm to improve classification rate to achieve greater accuracy.

Shajahan et al (2013) worked on the application of data mining techniques to model breast cancer data using decision trees to predict the presence of cancer. Data collected contained 699 instances (patient records) with 10 attributes and the output class as either benign or malignant. Input used contained sample code number, clump thickness, cell size and shape uniformity, cell growth and other results physical examination. The results of the supervised learning algorithm applied showed that the random tree algorithm had the highest accuracy of 100% and error rate of 0 while CART had the lowest accuracy with a value of 92.99% but naïve bayes' had the an accuracy of 97.42% with an error rate of 0.0258.

Mangasarian et al (1995) performed classification on both diagnostic and prognostic breast cancer data. The classification procedure adopted by them for diagnostic data is called Multi Surface Method-Tree (MSM-T) that uses a linear programming model to iteratively place a series of separating planes in the feature space of the examples. If the two sets of points are linearly separable, the first plane will be placed between them. If the sets are not linearly separable, MSM-T will construct a plane which minimizes the average distance of misclassified points to the plane, thus nearly minimizing the number of misclassified points. The procedure is recursively repeated. Moreover they have approached the prognostic data using Recurrence Surface Approximation (RSA) that uses linear programming to determine a linear combination of the input features which accurately predicts the Time-To-Recur (TTR) for a recurrent breast cancer case. The training separation and the prediction accuracy with the MSM-T approach was 97.3% and 97 % respectively whereas the RSA approach was able to give accurate prediction only for each individual patient. Their drawback was the inherent linearity of the predictive models.

Lundin et al (1999) has applied ANN on 951 instances dataset of Turku University Central Hospital and City Hospital of Turku. To evaluate the accuracy of neural networks in predicting 5, 10 and 15 years breast cancer specific survival. From the experiment the values of ROC curve for 5 years was evaluated as 0.909, for 10 years 0.086 and for 15 years 0.883, these values were used as a measure of accuracy of the prediction model. The author compared 82/300 false prediction of logistic regression with 49/300 of ANN for survival estimation and found ANN predicted survival with higher accuracy. It shows that neural networks are valuable tools in cancer survival prediction. In future the study should concentrate on collecting data from a more recent time period and find new potential prognostic factors to be included in a neural network model.

Delen et al (2005) compared ANN, decision tree and logistic regression techniques for breast cancer prediction analysis. They used the SEER data of twenty variables in the prediction models. From the experiment the author found that the decision tree with 93.6% accuracy and ANN with 91.2% are more superior to logistic regression with 89.2% accuracy. The study is based on multiple prediction models for breast cancer survivability using large datasets along with 10 fold cross validation method. It provides a relative prediction ability of different data mining methods. In future this work is extended by collecting real dataset in the clinical laboratory

2.1 Data Mining Process

Data mining is the process of extracting patterns from data; these patterns may be discovered depending on the data mining tasks that are applied on the dataset. The two basic data mining tasks are: descriptive data mining tasks which help to understand the characteristic properties of dataset and predictive data mining tasks which are used to perform predictions based on available dataset. Predictive data mining is the chosen data mining task for this study.

According to Gupta et al (2011) data mining applications can use different parameters to examine data which includes; association (patterns that define the relationship between data), sequence/pattern analysis (patterns where one event leads to another), classification (identification of new patterns with predefined targets) and clustering (grouping of identical of smaller objects). The basic steps include:

- *Problem definition* is the definition of the goals and objectives and the identification of tools to be used to build the defined model.
- *Data exploration* is the recommendation for useful dataset if the existing dataset does not meet the required need for analysis.
- *Data preparation* is the process of cleaning and transforming data to remove missing and invalid data and validation of data for robust analysis.
- *Modeling* is based on the desired outcomes and data. This involves the use of data mining algorithms (for this study; naïve bayes, decision trees and multi-layer perceptron) in meeting the necessary objectives-which for the purpose of this study is classification.
- *Evaluation and deployment* is the analysis and interpretation of the results of analysis to create recommendations for consideration.

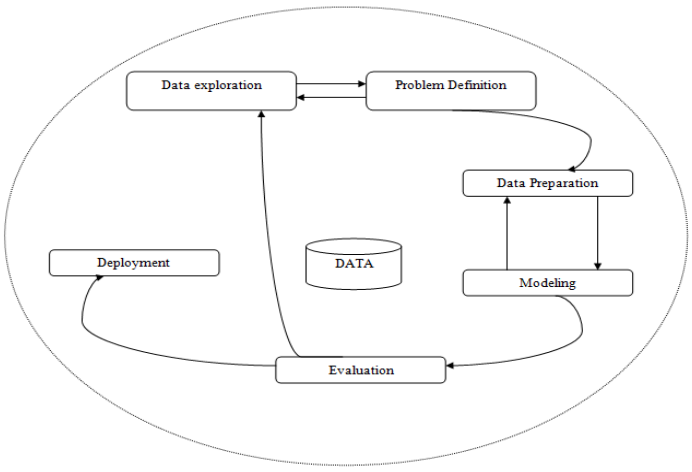


Figure 1: Data Mining Process Representation

3 Materials and Methods

In order to classify the breast cancer data collected form LASUTH with the aim of achieving high accuracy and precision; two supervised learning algorithms i.e., J48 decision trees and the naïve bayes are used. The data preprocessing was performed in order to remove inconsistent data and the data converted into a format that is useful in the simulation environment. WEKA data mining software was the environment used for simulating the breast cancer risk prediction model; which is an open-source data mining software used for academic purposes.

3.1 Training dataset description

LASUTH breast cancer data set was collected from the Cancer Registry of LASUTH, Ikeja in Lagos, Nigeria. The dataset collected contains 69 instances with 17 attributes. The class distribution is framed as unlikely, likely and benign. Hence there are 16 independent variables and 1 dependent variable. The nominal values are set for the independent variables and the dependent variable. The non-modifiable factors are the first 11 variables while the modifiable factors are the next 5 variables while the breast cancer risk is the last variable.

Table 1: The Training Dataset Description

S/N	Risk Factor (Points)	Values
1.	Family History of Breast Cancer	Yes, No
2.	Existence of Benign Breast disease	Never, Ever
3.	Mammographically Dense Breast	Never, Ever
4.	Age at First Birth	no, ≤30yrs, >30yrs
5.	Age at Menopause	no, ≤50yrs, >50yrs
6.	Body Mass Index (BMI)	< 25, ≥ 25
7.	Age at Menacre	no, ≤12yrs, >12 yrs
8.	Endogenous Estrogen Levels	Low, High
9.	Waist-Hip Ratio	< 0.81, ≥ 0.81
10.	Age	≤50yrs, >50yrs
11.	Sex	Male, Female
12.	Smoking Frequency	Never, former, current
13.	Alchohol Intake	Never, former, current
14.	Occupational Hazard	No, Yes
15.	Current Oral Contraceptive use	Never, Ever, Current
16.	Breast Feeding	Never, Ever
17.	Breast Cancer Risk	Unlikely, Likely, Benign

3.2 Data mining algorithms used

3.2.1 Naïve Bayes' classifier

Naive Bayes Classifier is a probabilistic model based on Baye's theorem. It is defined as a statistical classifier. It is one of the frequently used methods for supervised learning. It provides an efficient way of handling any number of attributes or classes which is purely based on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge on observed data. Let X is a data sample containing instances, X_i where each instances are the breast cancer risk factors (modifiable and non-modifiable). Let H be a hypothesis that X belongs to class C which contains (unlikely, likely and benign cases). Classification is to determine $P(H_j|X)$, (i.e., posteriori probability): the probability that the hypothesis, H_j (unlikely, benign or likely) holds given the observed data sample X.

- $P(H_j)$ (prior probability): the initial probability of the hypothesis in the class;
- $P(X_i)$: probability that sample data is observed for each attribute, i;
- $P(X_i|H)$ (likelihood): the probability of observing the sample's attribute, X_i given that the hypothesis holds in the training data X; and
- posteriori probability of a hypothesis H_j (unlikely, likely or benign), $P(H_j|X_i)$, follows the Baye's theorem as follows:

For example, if for a variable X with i attributes (breast cancer risk factors) expressed as:

$X = \{X_1, X_2, X_3, X_4, \ldots\ldots\ldots, X_{1}\}$ and

H_j={unlikely, likely, benign}.

Then,

is the probability of the outcome of a risk factor being under the hypothesis, H_j;

is the probability of the outcome of the risk factor in the training dataset;

is the probability of the outcome of an hypothesis (unlikely, likely, benign i.e. j=3);

is the probability of a variable, X containing risk factors belongs to an hypothesis, H_j;

The breast cancer risk class output = maximum $[P(H_j|X)]$ for j=1, 2, 3

3.2.2 Decision Trees

J48 decision trees classifier is a simple decision learning algorithm, it accepts only categorical data for building a model. The basic idea of ID3 is to construct a decision tree by employing a top down greedy search through the given sets of training data to test each attribute at every node. It uses statistical property known as information gain to select which attribute to test at each node in the tree. Information gain measures how well a given attribute separates the training samples according to their classification.

It is suitable for handling both categorical as well as continuous data. A threshold value is fixed such that all the values above the threshold are not taken into consideration. The initial step is to calculate information gain for each attribute. The attribute with the maximum gain will be preferred as the root node for the decision tree.

Given a set S of breast cancer cases, J48 first grows an initial tree using the divide-and-conquer algorithm as follows:

- If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S;
- Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S_1, S_2,……, S_n for a dataset containing n cases according to the outcome for each case, and apply the same procedure recursively to each subset.

It uses a statistical property known as information gain to select which attribute to test at each node in the tree. It measures how well a given attribute separates the training samples according to their classification.

3.2.3 Performance Evaluation

The performance evaluation criteria allow the measurement of the accuracy of the models developed using the training dataset. The results of the classification are recorded on a confusion matrix. A confusion matrix is a square which shows the actual classification along the vertical and the predicted along the vertical. All correct classifications lie along the diagonal from the north-west corner to the south-east corner also called True Positives (TP) and True Negatives (TN) while other cells are called the False Positives (FP) and False Negatives (FN). If the unlikely case is considered positive then likely and benign are called negatives, if likely is considered as positive then unlikely and benign are considered negatives and the same also applies if benign is called the positive. These values are used to determine the following evaluation criteria.

The error rates of the developed models using both classifiers were also determined alongside with the performance evaluation criteria mentioned above.

4 Experimental Results and Discussions

The experimental results of this study using the two classifiers are discussed using the WEKA software data mining tool. As earlier discussed, breast cancer is classified as either unlikely, likely and benign. The performance evaluation results and the error rates are also discussed as follows.

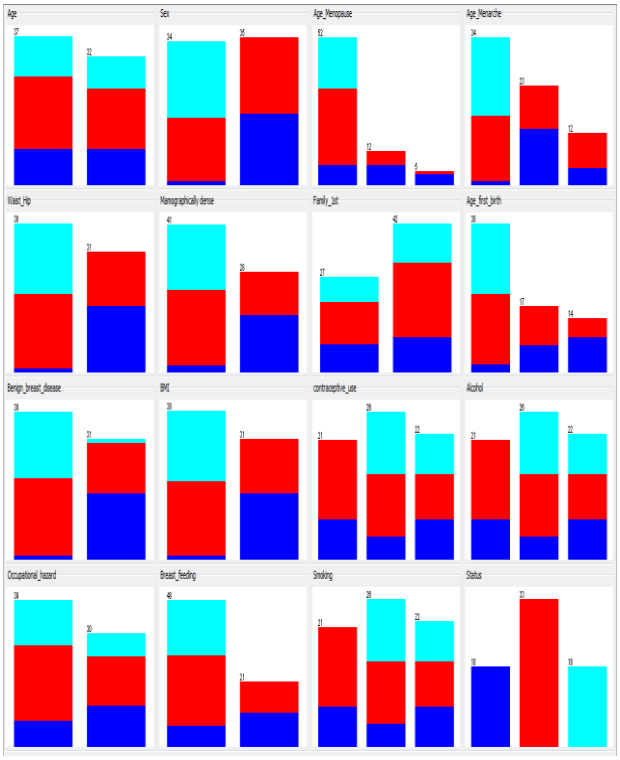


Figure 2: Distribution of the training dataset used

a	b	c	<- - classified
17	1	0	a = benign
6	22	5	b = Likely
0	0	18	c = Unlikely

a	b	c	<- - classified
17	1	0	a = benign
2	31	0	b = Likely
0	1	17	c = Unlikely

Figure 3: Confusion matrix of the results of classification using J48 decision trees (left) and naïve bayes' (right)

From the results of the data mining process for the prediction of Breast Cancer risk using J48 decision trees and Naïve Bayes' classifiers, the confusion matrix of both models can be seen in Figure 3.

The results of the J48 decision trees showed that from the 69 training data collected, out of 18 cases that were benign 17 were correctly classified and 1 incorrectly classified as Likely; out of the 33 cases that were likely 22 were correctly classified with 6 and 5 incorrectly classified as benign and unlikely respectively and form the total 18 cases that were unlikely all were correctly classified with no misclassifications.

The results of the naïve bayes' classifier showed that out of the 69 training data collected; out of 18 cases that were benign 17 were correctly classified with 1 misclassified as likely; out of the 33 cases

that were likely 31 were correctly classified with 2 misclassified as benign and out of the 18 cases that were unlikely 17 were correctly classified and 1 misclassified as likely.

From the two confusion matrices, it can be seen that the naïve bayes' model had 57 correct and 12 incorrect classifications giving an accuracy of 82.6% but the J48 decision trees which had correct and incorrect classifications of 65 and 4 respectively had an accuracy of 94.2% (see Table 2 and Figure 4 below).

Table 2: Accuracy of naïve bayes' and J48 decision trees' model

	Naïve Bayes'	J48 Decision Trees
Correct Classification	57	65
Incorrect Classification	12	4
Accuracy (%)	82.6	94.2

From the two models developed for the prediction of breast cancer risk; the confusion matrix developed earlier was used to identify the accuracy of the models; other performance evaluation criteria are as follows. The True Positive (TP) rate/recall which is the percentage of the actual number of positive that were classified as positive cases has an average of 87% and 94% for the naïve bayes' and decision trees respectively. The False Positive (FP) rate which is the percentage actual number of positive cases that were misclassified also called false alarm has an average of 8.1% and 3.1% for the naïve bayes' and decision trees respectively. These results of the TP and FP rate have a value of 96.7% and 99% for the area under the graph of the Receiving Operating Characteristics (ROC) for naïve bayes' and decision trees respectively; this is a good indication of the effectiveness of both models but with the values of the TP rate, FP rate, Area under the ROC and accuracy of the models; the decision tree is a better model with an average precision of 94.4% compared with 82.6% for the naïve bayes' model (see Table 3 below). The error rates of the two models are 0.1396 and 0.058 for the mean absolute error and 0.3242 and 0.1703 for the relative absolute error of the naïve bayes' and the J48 decision trees model respectively (see Figure 5 below).

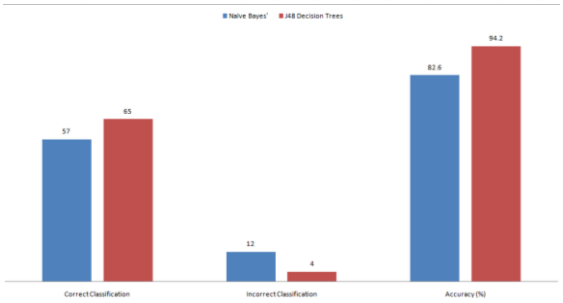


Figure 4: Accuracy, Correct and incorrect classification of Breast Cancer by both models

Table 3: Performance evaluation of both models

Performance criteria	Naïve Bayes'				J48 Decision Trees			
	Unlikely	Likely	Benign	Average	Unlikely	Likely	Benign	Average
TP rate	1	0.667	0.944	**0.870333**	0.944	0.939	0.944	**0.942333**
FP rate	0.098	0.028	0.118	**0.081333**	0	0.056	0.039	**0.031667**
Precision	0.783	0.957	0.739	**0.826333**	1	0.939	0.895	**0.944667**
ROC Area	0.995	0.953	0.953	**0.967**	0.998	0.985	0.987	**0.99**

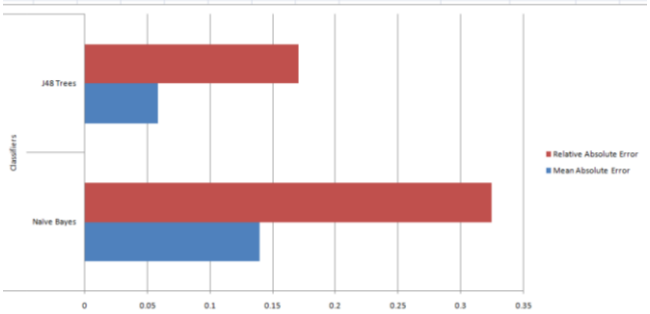


Figure 5: Error rate for both model

Figure 6 below gives an expression of the rules developed by the J48 decision trees model for the prediction of breast cancer risk using the dataset collected for cancer patients of LASUTH. It gives a clear picture of understanding better the relationship between each attributes and breast cancer risk.

If Waist –Hip ratio = <0.81
- If Contraceptive Use = Current then Breast Cancer = Likely
- If Contraceptive Use = Never
  - If Sex = Male then Breast Cancer = Unlikely
  - If Sex = Female then Breast Cancer = Likely
- If Contraceptive Use = Ever
  - If Benign Breast Disease = Never
    - If Occupational Hazard = No then Breast Cancer = Unlikely
    - If Occupational Hazard = Yes then Breast Cancer = Likely
  - If Benign Breast Disease = Ever then Breast Cancer = Likely

If Waist – Hip ratio = >=0.81
- If Benign Breast Disease = Never then Breast Cancer = Likely
- If Benign Breast Disease = Ever
  - If Contraceptive Use = Current then Breast Cancer = Benign
  - If Contraceptive Use = Never
    - If Occupational Hazard = No then Breast Cancer = Likely
    - If Occupational Hazard = Yes then Breast Cancer = Benign
  - If Contraceptive Use = Ever then Breast Cancer = Benign

Figure 6:Rules created from the dataset using J48 decision trees for predicting cancer risk

From the above results shown, it is very clear that data mining techniques can be used in predicting breast cancer risks and that the J48 decision trees has a better accuracy than the naïve bayes' model which is a statistical tool. This is the rule that was used by the decision trees in testing the model using the test data and the decision trees shows that the best attributes for predicting breast cancer are: Waist-Hip ratio, Contraceptive use, Sex, Benign breast disease and Occupational hazard.

5 Conclusion

In this study two different data mining classification techniques was used for the prediction of breast cancer risk and their performance was compared in order to evaluate the best classifier. Experimental results shows that the J48 decision trees is a better model for the prediction of breast cancer risks for the values of accuracy, recall, precision and error rates recorded for both models. Hence, an efficient and effective classifier for breast cancer risks has been identified while the number of attribute covered by the classifier can be increased by increasing the sample size of the training set and hence the development of a more accurate model.

REFERENCES

[1] American Cancer Society (2005). "Breast Cancer Facts & Figures 2005–2006" (PDF). Archived from the original on 13 June 2007. http://web.archive.org/web/20070613192148/http://www.cancer.org/downloads/STT/CAFF2005BrFacspdf2005.pdf. Retrieved 2013-02-26.

[2] American Cancer Society (2007). "Cancer Facts & Figures 2007" (PDF). Archived from the original on 10 April 2007. http://web.archive.org/web/20070410025934/http://www.cancer.org/downloads/STT/CAFF2007PWSecured.pdf. Retrieved 2012-11-26.

[3] American Cancer Society (2007). "Cancer Facts & Figures 2007" (PDF). Archived from the original on 10 April 2007. http://web.archive.org/web/20070410025934/http://www.cancer.org/downloads/STT/CAFF2007PWSecured.pdf.

[4] Blackburn, GL; Wang, KA (2007). "Dietary fat reduction and breast cancer outcome: results from the Women's Intervention Nutrition Study (WINS)." *The American journal of clinical nutrition* **86** (3): s878-81. PMID 18265482.

[5] Boffetta P, Hashibe M, La Vecchia C, Zatonski W, Rehm J (August 2006). "The burden of cancer attributable to alcohol drinking". *International Journal of Cancer* **119** (4): 884–7. doi:10.1002/ijc.21903. PMID 16557583.

[6] Boris Pasche (2010). *Cancer Genetics (Cancer Treatment and Research)*. Berlin: Springer. pp. 19–20. ISBN 1-4419-6032-5.

[7] Collaborative Group on Hormonal Factors in Breast Cancer (August 2002). "Breast cancer and breastfeeding". *Lancet* **360** (9328): 187–95. doi:10.1016/S0140-6736(02)09454-0. PMID 12133652.

[8] Delen, D., Walker, G., Kadam, A. (2005) Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, vol. 34, pp. 113-127, June 2005.

[9] Ferro, Roberto (1 January 2012). "Pesticides and Breast Cancer". *Advances in Breast Cancer Research* **01** (03): 30–35. doi:10.4236/abcr.2012.13005.

[10] Gage, M; Wattendorf, D; Henry, LR (1 April 2012). "Translational advances regarding hereditary breast cancer syndromes". *Journal of surgical oncology* **105** (5): 444–51. doi:10.1002/jso.21856. PMID 22441895.

[11] Grey, N and Sener, S. (2006) Reducing the global cancer burden, http://www.hospitalmanagement.net/features/feature648/ , Date accessed 21 November 2012.

[12] Gupta, S.; Kumar, D., Sharma, A (2011). Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis. Indian Journal of Computer Science and Engineering (IJCSE). Vol. 2 No. 2 pg 198-195, April, 2011. ISSN: 0976-5166.. Accessed on June 24, 2014.

[13] Hendrick, RE (October 2010). "Radiation doses and cancer risks from breast imaging studies.". *Radiology* **257** (1): 246–53. doi:10.1148/radiol.10100570. PMID 20736332

[14] Johnson KC, Miller, AB, Collishaw, NE, Palmer, JR, Hammond, SK, Salmon, AG, Cantor, KP, Miller, MD, Boyd, NF, Millar, J, Turcotte, F (2009). "Active smoking and secondhand smoke increase breast cancer risk: the report of the Canadian Expert Panel on Tobacco Smoke and Breast Cancer Risk (2009).". *Tobacco control* **20** (1): e2. doi:10.1136/tc.2010.035931. PMID 21148114.

[15] Lundin M., Lundin J., BurkeB.H.,Toikkanen S., Pylkkänen L. and Joensuu H.,(1999) "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer", Oncology International Journal for Cancer Resaerch and Treatment, vol. 57, 1999.

[16] Mangasarian,D.S.;Street, W.N.,Wolberg, W.H (1995). Breast cancer diagnosis and prognosis via linear programming, Operations Research, 43(4), pages 570-577, July-August 1995.

[17] Parkin, D.M., Ferlay, J., Hamdi-Cherif, M., Sitas, F., Thomas, J.O., Wabinga, H., Whelan, S.L. (2003). Cancer in Africa Epidemiology and Prevention, IARC (WHO) Scientific Publications no. 153, IARC Press, Lyon, France.

[18] Poongodi, M., Manjula, L., Pradeepkumar, S., Umadevi, M. *Cancer Prediction Technique Using Fuzzy Logic.* International Journal of Current Research, Vol. 3, Issue 11, pg 333-336, December 12, 2001. http://www.journalera.com. ISSN: 0975-833X. Accessed on June 24, 2014.

[19] Rajesh, K., Anand, S (2012). Analysis of SEER dataset for breast cancer diagnosis using C4.5 classification algorithm. International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 2, April 2012. ISSN 2278-1021. http://www.ijarcee.com pg. 72 – 77.

[20] Santoro, E., DeSoto, M., and Hong Lee, J (February 2009). "Hormone Therapy and Menopause". National Research Center for Women & Families. http://www.center4research.org/2010/03/hormone-therapy-and-menopause/.

[21] Sariego J (2010). "Breast cancer in the young patient". *The American surgeon* **76** (12): 1397–1401. PMID 21265355.

[22] Shajahaan, S.S; Shanthi, S., Chitra, V.M. (2013). Application of Data Mining Techniques to model Breast Cancer Data. International Journal of Emerging Technology and Advanced Engineering Vol 3, Issue 11, November 2013. ISSN 2250-2459. http://www.ijetac.com pg 362 – 369.

[23] WHO. (2002). National Cancer Control Programmes; policies and managerial guidelines, 2nd edition.

[24] Yager JD (2006). "Estrogen carcinogenesis in breast cancer". *New Engl J Med* **354** (3): 270–82. doi:10.1056/NEJMra050776. PMID 16421368.

TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

Applying Fuzzy Cluster Index to Improve Searching in Data Warehouse

[1]Ahmad A. Almafrji, [2]Shawkat K. Guirguis and [3]Magda M. Madbouly
Department of Information Technology, Institute of Graduate Studies & Research, Alexandria University; Alexandria, Egypt
[1]ahabm2012@yahoo.com; [2]shawkat_g@yahoo.com; [3]mmadbouly@hotmail.com

ABSTRACT

Data Warehouse (DW) is one of the solutions for decision-making process in a business organization. But it only stores data for managerial purpose and it has no intelligent mechanism for decision making. For improving the process of decision making and searching Data Warehouse (DW) of the medical resources (items), where this study includes an application on a Data Warehouse (DW) of medical resources (items).In this paper, we merged the fuzzy rule with cluster index technique. where The proposed technique is named Fuzzy Cluster Index technique (FCI) to improve and speed up Queries fuzzy rule and process of decision making and management medical (items), The performance evaluation of three data warehouse queries is focused in this paper by comparing with Fuzzy cluster index technique (FCI), Fuzzy Rule and Index-based Apriori Algorithm to observe the results of variable size dataset with respect to time. Eventually, the designed system was constructed and executed by using (C# version 2010) which is a visual and object oriented programming language. This proves the efficiency of the proposed system for improving searching in Data Warehouse (DW) and the decision support system for the medical items in a perfect way.

Keywords: Data Warehouse, Apriori Algorithm, Fuzzy Cluster Index (FCI), Fuzzy Rule.

1 Introduction

Almost every enterprise uses a database to store its vital data and information For instance dynamic websites, accounting information systems payroll systems, stock management systems all rely on internal databases as a container to store and manage their data. In fact, data warehousing is the process of collecting data from operational functional databases, transforming, and then archiving them into special data repository called data warehouse with the goal of producing accurate and timely management information. [1]. A Data Warehouse (DW) is defined as "a subject-oriented, integrated, time-variant non-volatile collection of data in support of management's decision-making process" Data warehouses store huge amount of information from multiple data sources which is used for query and analysis, the data is stored in the multidimensional (MD) structure Multidimensional modeling requires specialized design techniques that resemble the traditional database design methods .[2].

1.1 Data Warehouse Models

The data model mainly used for database designing is the Entity/Relationship model (ERM). This type of model, however, presents problems: as in reality, entities have different characteristics, contain a

DOI: 10.14738/tnc.32.1021
Publication Date: 13th March 2015
URL: http://dx.doi.org/10.14738/tnc.32.1021

different quantity of data, etc. Therefore it is necessary to adopt a multi-dimensional view. To allow a multi-dimensional visualization of data, techniques have been developed known as “schema”. [3].

- Star schema.
- Snowflake schema.
- Mixed schema.

1.2 Fuzzy Data Warehouse

The numeric values of a classical data warehouse can be difficult to understand for business users, or may be interpreted incorrectly, Therefore for a more accurate interpretation of numeric values, business users require an interpretation in meaningful non-numeric terms fuzzy data warehouse which allows integration of fuzzy concepts directly into the data warehouse By using this approach, the concept of summarize ability is not affected in dimensions as the fuzzy concepts are rolled out in a meta-table structure.The proposed approach is more flexible as it allow integrating and redefining fuzzy concepts without the need for redesigning the core of a data warehouse.[4].

The theory of fuzzy sets facilitates the coding of human knowledge in the form of linguistic concepts. For example, the concept product promotion impact can be scored as low, and high. Each promotion can be assigned degree values (usually between 0 and 1) for each label (e.g., 0.3/low, 0.9/high). Afterwards, these values can be incorporated into a computational framework that will support a decision process (e.g., approval of a promotion).When important business data or business measures or entities are fuzzy, it may be useful to construct a fuzzy data warehouse that can directly support the analysis of fuzzy data.

1.3 Comparison of Classical and Fuzzy Data Warehouse

The following table presents a comparison of the classical and the fuzzy data warehouse model in order to summarize the main advantages of the fuzzy data warehouse approach. [5].

Classical Data Warehouse	Fuzzy Data Warehouse
In a classical DWH, an instance does not belong to more than one class at a time. Because of this, true values of the classification cannot be measured.	Classification of dimension attributes or facts in the FDWH is done in a fuzzy manner, allowing values to belong to more than one class and the classification to be more accurate.
Qualitative interpretation of facts and dimension attributes is not supported in a classical DWH.	A FDWH enables using non-numerical attributes. As a result, both qualitative and quantitative attributes can be used for analysis
Decision-making processes are often verbal. A classical DWH approach does not include any linguistic concept to interpret the data.	The definition of linguistic variables can be derived from the business environment manually. This reduces the effort of interpreting numeric values and facilitates decision-making processes.
Only crisp data is used for analysis and decision making.	Both fuzzy and crisp data can be used for analysis and decision making.
The classical schema consists of dimensions and facts.	The FDWH schema consists of a classical schema together with fuzzy meta tables called fuzzy classification table and fuzzy membership table.
In a classical DWH only extracted data (slices, dices, etc.) can be classified. The classification can therefore not be propagated on other hierarchy levels of dimensions.	In FDWH the fuzzy concepts can be propagated over the dimensions in order to apply the classifications on other hierarchy levels.
The Retrieval of queries in a classical DWH is based on SQL in most cases.	A FDWH can be queried on a linguistic level. For example, fCQL (Meier et al.) allows marketers to classify single customers or customer groups by classification predicates such as ‘loyalty is high and turnover is large

2 Related Work

Lately data warehousing (DW) has gained a lot of attention both from both the industry and research community communities. From the industrial perspective, building an information system for the huge data volumes in any industry requires lots of resources as time and money. Unless those resources add to the industry value, such systems are worthless. Thus, people require that information systems should be capable to provide extremely fast responses to different queries specially those queries that affect decision making. Data warehousing systems address the issue of enabling managers to acquire and integrate information from different sources, and to efficiently query very large databases. The best decisions are made when all the relevant data is taken into consideration. Today, the biggest challenge in any organization is to achieve better performance with least cost, and to make better decisions than competitors. That is why data warehouses are widely used within the largest and most complex businesses in the world. data warehouse (DW) is a collection of consistent, subject-oriented, integrated, time-variant, nonvolatile data along with processes on them, which are based on current and historical information that enable people to make decisions and predictions about the future. The DW is suitable for direct querying and analysis, and it stands as a source for building logical data marts oriented to specific areas of an enter prize.[12]

The authors in [6] presented find out performance optimization and Enhancement techniques which improve the processing time and faster data retrieval in data warehousing .We have seen Performance Optimization and Enhancement Techniques of Data Warehouse. Data Warehousing is not a new phenomenon. All large organizations already have data warehouses, but they had some difficulty to managing them properly. Data warehouse Performance is heavily dependent on proper indexing strategy. B-Tree indexes and bitmapped indexes are suitable. A proper indexing technique is crucial to avoid I/O intensive table scans against large data warehouse tables. It also depends upon the System Configuration and Volume of Data. So proper selecting of right techniques for storing as well as retrieving is necessary for data warehouse other performance improvement schemes that are part of the physical design include the following: data partitioning, data clustering, parallel processing, creation of summaries, adjusting referential integrity checks, proper setting of DBMS Initialization Parameters and Use of Data Array.

The authors in [7] presented this research is to compare some data models considering their data density and their data sparsely management to optimize Data Warehouse environments. In this research paper various techniques for query performance optimization have been explored and a close association of its conceptual aspects with Oracle Warehouse Builder is mapped.

The authors in [9] presented offers analysis and comparison of some of the related facts, which have been drawn from past resources that concern on bitmap indexing for data warehouses. Those resources are reviewed in this research one by one. Due to the importance of DW which is an important element for BI and the main role to improve strategic decision making, this technology is discussed in this research. Moreover, there are many popular techniques are used to enhance the DW treatment queries performance such as indices, views of materialized, and fragmentation of data. The aim of this research is to analyze and to compare many related techniques that concern on bitmap indexing for data warehouses.

The authors in [10] presented new approach of data warehouse minimization by fuzzy-based ETL filter for ETL processes in business intelligence (BI) systems. First part introduces common company

systems and possible data sources in the company. Second part states the problem with interpreting information in BI systems and explains a data representation in the BI systems. Third part of the paper identifies suitable linguistic variables that help with interpreting the data to the user and automated filter as well. We also define a rule base and input and output values of expert system. Last part of the paper proposes a two ways to minimize a data - modification border of the fuzzy set and omitting useless combinations of the linguistic variables and modifiers.

The authors in [11] presented a model of a web-based system for knowledge warehousing and mining of diagnosis and therapy of HIV/AIDs using Fuzzy Logic and data mining approach. A model was developed, using the predictive modeling technique, for predicting HIV/AIDs and monitoring of patient health status. The fuzzy inference rule and a decision support system based on cognitive filtering was employed to determine the possible course of action to be taken. A case study of some data of PLWH was used and the result obtained shows that the developed system is efficient. The system uses XAMP on Windows OS platform. The system was tested and evaluated with satisfactory results.

The authors in [13] presented in some steps, a comparative study between the index B-tree and Bitmap type, their advantages and disadvantages, with a real experiment based on two factors: size of index and clustering factor, this shows that the Bitmap index is more advantageous than the B-tree one. By using the B-tree index, the optimizer opted for a full table scan; this operation makes a higher clustering factor, whereas in the case of bitmap index that makes a low Clustering factor, he used to answer the query. You can deduct the performance by the number of I / O required fetching the result.

3 Proposed Methodology

Characteristics of Proposed System: our Proposed system used to improve and speed up Queries fuzzy rule and process of decision making and management medical (items), i.e., we merged the fuzzy rule with cluster index technique. Where the proposed technique is named Fuzzy Cluster Index technique (FCI), the performance evaluation of three data warehouse queries is focused in this paper by comparing with Fuzzy cluster index (FCI), Fuzzy Rule and Index-based Apriori Algorithm to observe the results of variable size dataset with respect to time. Figure 1: Shows our proposed System.

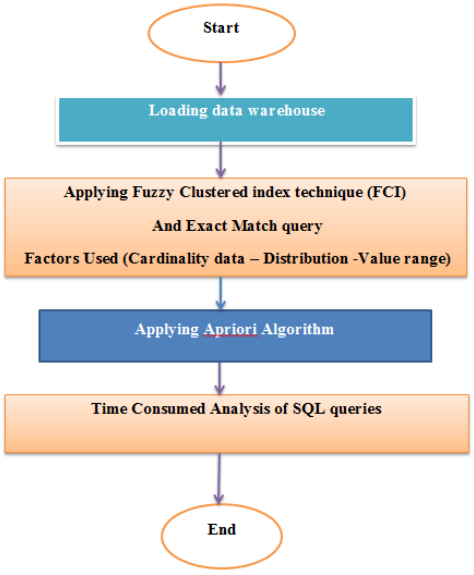


Figure 1: Our proposed System

The Potential problems: In this section, will explain the potential problems that facing of data warehouse users, and the most important cases that may be exposed to data warehouse users, and proposed solution for this cases:

- Using large amounts of data in applications data warehouse and the search process takes a great time, lead to not good decisions by decision-makers.
- Data warehouses are used for analysis of businesses performance. Potential pitfall of the classical data warehouse is that the numeric values of a data warehouse may be difficult to interpret for business users, or may be interpreted incorrectly. For more accurate understanding of numeric values, business users require an interpretation in meaningful, non-numeric terms.

The proposed technical to solve the problem:

in this work a specific system applying fuzzy logic to improve the search in the data warehouse and to reduce the time of the search in a data warehouse that will help decision makers to take the decision easily and more flexible because the fuzzy logic is based on natural language.

3.1 System components:

3.1.1 Apriori Algorithm:

Apriori is very much basic algorithm of Association rule mining. It was initially proposed by R. Agrawal and R Srikant for mining frequent item sets. This algorithm uses prior knowledge of frequent item set properties that is why it is named as Apriori algorithm. Apriori makes use of an iterative approach known as breath-first search, where k-1 item set are used to search k item sets. There are two main steps in Apriori. 1) Join - The candidates are generated by joining among the frequent item sets level-wise. 2) Prune- Discard items set if support is less than minimum threshold value and discard the item set if its subset is not frequent [15].

3.1.2 How Index-based Apriori Algorithm Work

The work steps of Index-based Apriori algorithm are explained in detail with examples. These steps are divided to two main parts which are:

1) Steps of generating candidate and supported item sets.

2) Steps of generating association rules from supported item sets.

1) Steps of Generating Candidate and Supported Item sets

The work steps of the first part of the Index-based Apriori algorithm for generating candidate and supported item sets are explained in detail through example, and figure shows the work mechanism of these steps. [8].

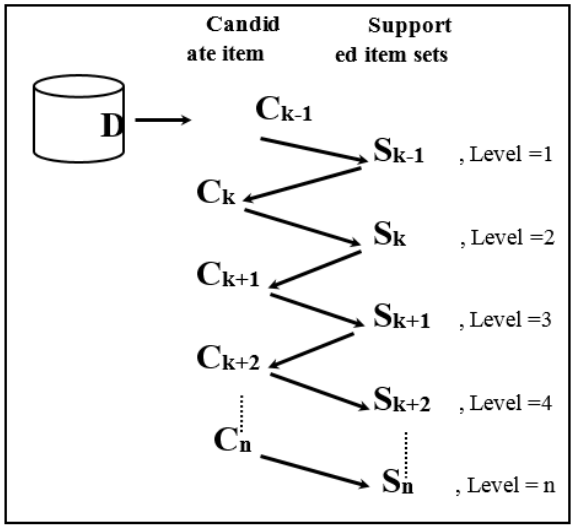


Figure 2: Diagram illustrating the Index-based Apriori algorithm mechanism for finding large set of supported item sets [8]

D is the database of items.
Ck is the set of candidate itemsets (supported and unsupported itemsets)
Sk is the set of supported itemsets
k is the number of level

Figure 3 demonstrates the steps (levels) of Index-based Apriori algorithm generating for candidate and supported item sets, where (*) refers to the join process and (**) Refer to the pruning method supposing that (mincount = 3), where the counts of Item sets are calculated by applying Indexing technique:

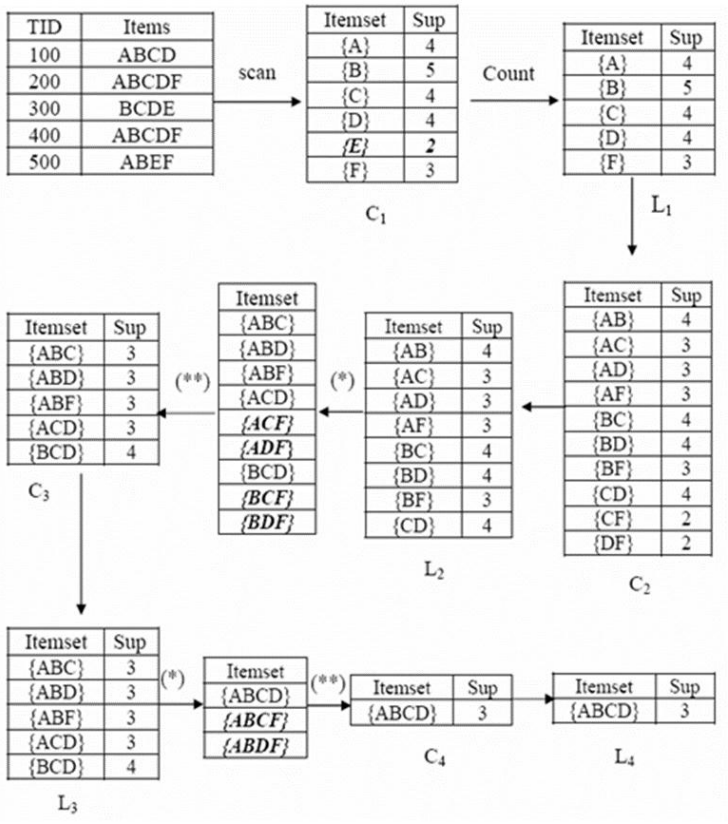


Figure 3: The Steps of Generating Candidate and Supported Item sets [8]

3.2 Fuzzy logic system

The basic configuration of a fuzzy logic system with fuzzifier and defuzzifier. This type of fuzzy logic system was first proposed by Mamdani. The main fours components' functions are shown in figure 4 below. [14].

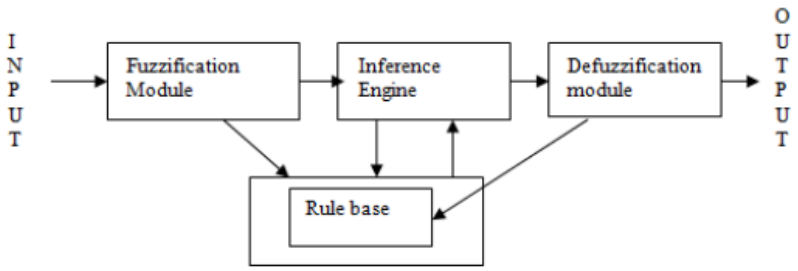


Figure 4: Fuzzy logic system

Fuzzification module: This module does a mapping from input to a fuzzy set.

Fuzzy Rule Base: Fuzzy logic systems use fuzzy IF-THEN rules. In a fuzzy logic system, the collection of fuzzy IFTHEN rules is stored in the fuzzy rule base which is referred to by the inference engine when processing inputs.

Fuzzy Inference Engine: previously all input values have been fuzzified into their respective linguistic values; the inference engine will access the fuzzy rule base of the fuzzy expert system to derive linguistic values for the intermediate as well as the output linguistic variables. The two main steps in the inference process are aggregation and composition. Aggregation is the process of computing for the values of the IF (antecedent) part of the rules while composition is the process of computing for the values of the THEN (consequent) part of the rules.

Defuzzification module: Defuzzifier does a mapping from the fuzzy output to the crisp output.

3.3 How Clustered Index Work

It is a type of index in which the data is arranged in distinct order (in sequence) which means clustered index determines the physical order of data in table. It is beneficial when there is need to access the records sequentially or in the reverse order. There can only be one clustered index per table, because the data rows themselves can only be sorted in one order. There are row locators which is clustered index key on the row. The only time the data rows in a table are stored in sorted order is when the table contains a clustered index. If a table has no clustered index, its data rows are stored in a heap. [16].

Example:

When creating a clustered index on First Name column, the data in the table is physically alphabetically sorted based on First Name value. When inserting a new row into the database, it will be inserted in a certain position so that the sorting is still kept.

1. The leaf nodes represent the actual data pages while the intermediate nodes of the tree structure are index pages. All the pages in the structure are linked.
2. The top node in the structure is the root index page, while the middle level nodes are intermediate index pages.
3. Each row in an index page refers either another index page or a data page. This reference is a pointer to a page number.

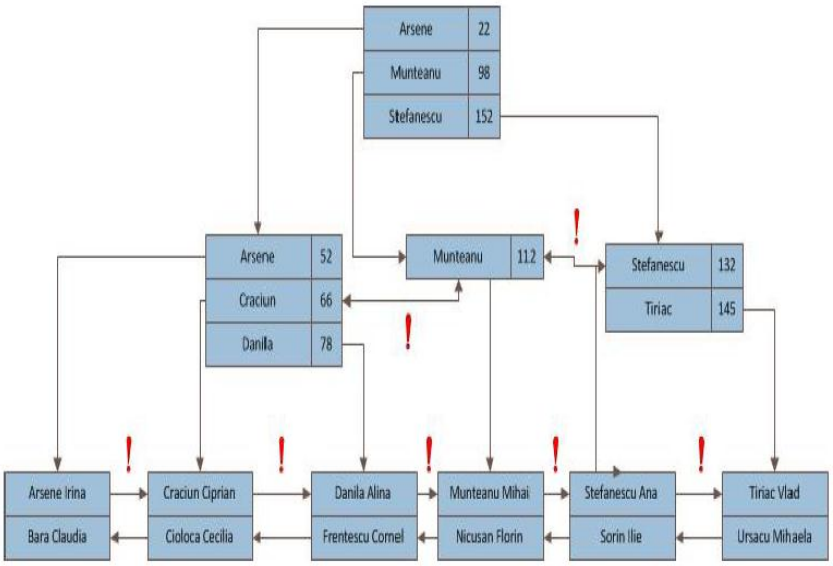


Figure 5: The structure of the clustered index [17]

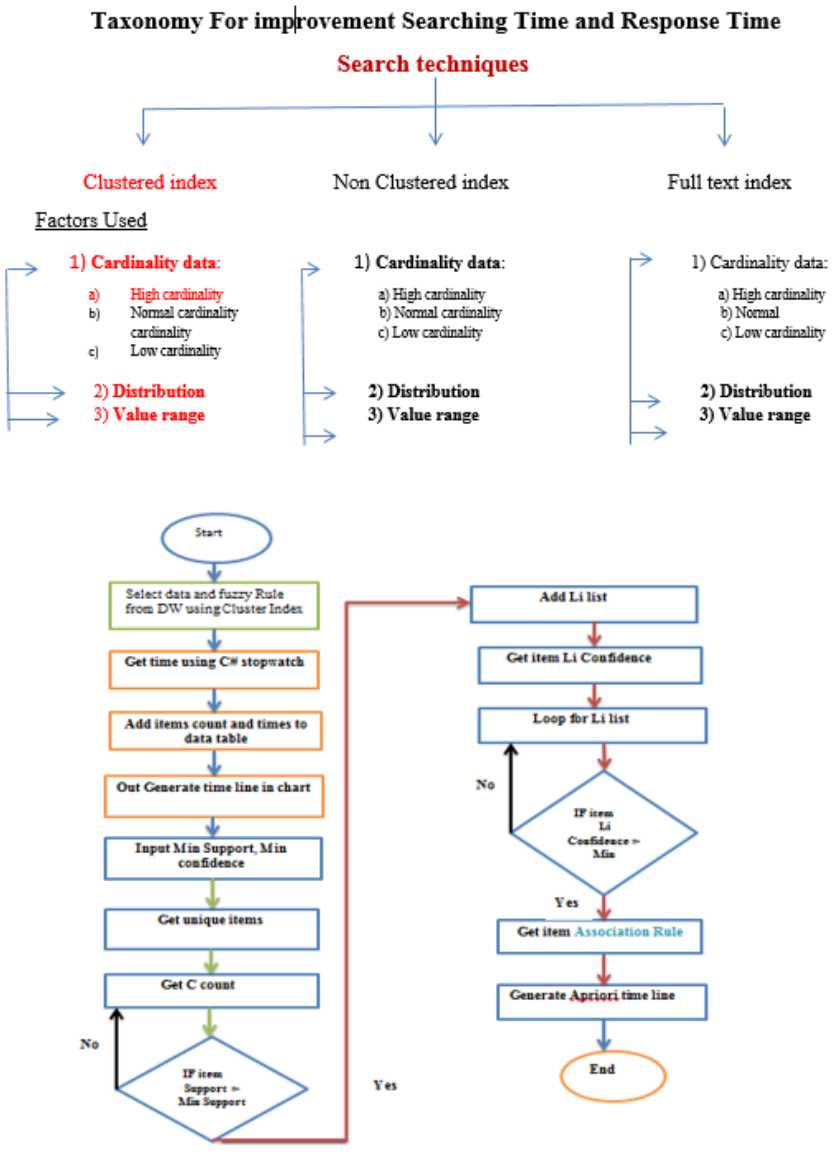


Figure 6: The proposed System architecture to improve searching in DW

Data flow in system architecture:

There are three scenarios for the data flow in the system:

- The first scenario begins the process of recording selected data and Fuzzy Rule from Data Warehouse and using Cluster Index to speed up Queries Fuzzy Rule.
- The second scenario begins the process execution Apriori Algorithm and on the same size of data that has been execution technique Fuzzy Rule Cluster index (FCI) them.
- The third scenario begins the Process execution Fuzzy Rule without Cluster Index and on the same size of data that has been execution first scenario and second scenario.

The results comparison between the three scenarios to demonstrate the technique evaluation of the proposed Fuzzy Rule cluster Index in terms of search time in the selected data from the Data Warehouse.

4 Experiments and Results

4.1 Research Material

4.1.1 Used Device:

The algorithm, presented in this thesis, is implemented with a laptop DELL model with the following specifications of Intel® Core(TM)i5 2430 M CPU @ 2.40 GHz 2.40 GHz with 8 GB RAM, system type 64-bit operating system. This machine is equipped with operating system Windows 7 Ultimate.

4.2 Research Results:

4.2.1 First experiment:

In the first experiment, shows from Table 1 and Figure 7 in SQL,"WHERE. Key word can use for searching of exact keyword in tables of data set. The Index-based Apriori Algorithm consumption is increasing, Fuzzy Rule time consumption is much less. But Fuzzy Cluster index technique (FCI) takes less time. Fuzzy Cluster index (FCI) gives better searching in exact matching of string and results good performance.

Table 1: comparison between the three techniques

Items	Fuzzy Cluster Index (FCI) Time/ S	Fuzzy Rule without Cluster Index Time/ S	Apriori Algorithm Time/ S	Mini Supp	Mini Conf
393665	7.0277	10.0598	87.3459	10	5
207199	5.9843	8.1338	46.7846	10	0
144808	3.3231	7.5228	42.4740	5	5
41658	0.5991	6.3775	12.3150	3	0

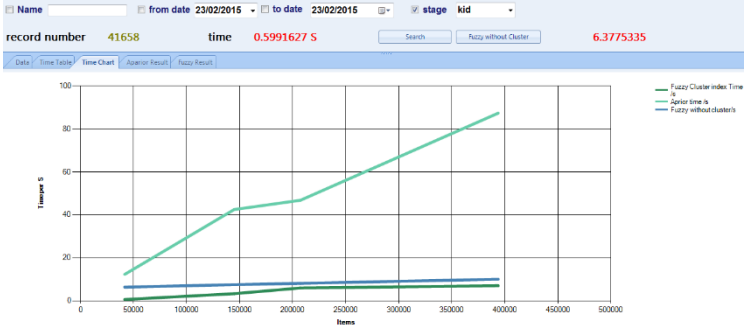


Figure 7: comparison between the three techniques

4.2.2 Second experiment:

In this Second experiment , we show table 2, Figure 8 The Index-based Apriori Algorithm time consumption is 232.7087 s, because the manager chose in Index-based Apriori Algorithm for minimum support= 2 and minimum confident=2 leads to a large search time and the loss of a small number of data and frequent large for data. fuzzy rule the time consumption is much less nearly 220 s approx. Fuzzy Cluster index technique (FCI) takes less time for searching even the records in Data Warehouse are increasing. The Fuzzy Cluster Index technique (FCI) and Fuzzy Rule having the equal elapsed time approx. but Index-based Apriori Algorithm elapsed time having the much elapsed time as the records are increasing from 41658 item to 393665 item.

Table 2: comparison between the three techniques

Items	Fuzzy Cluster Index (FCI) Time/ S	Fuzzy Rule without Cluster Index Time/ S	Apriori Algorithm Time/ S	Mini Supp	Mini Conf
393665	7.4687	10.4023	232.7087	2	2
207199	6.0426	8.003	61.3984	2	0
144808	3.8084	7.4670	42.8171	2	2
41658	0.5655	6.4057	13.1205	1	0

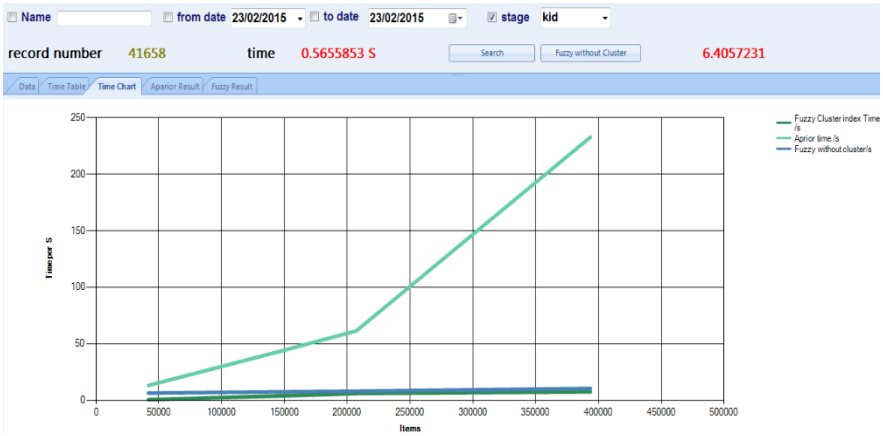


Figure 8: comparison between the three techniques

4.2.3 Third experiment:

In this third experiment, we show table 3, Figure 9 The Index-based Apriori Algorithm time consumption is 24.6466 s, the Index-based Apriori Algorithm searching time better in results good performance of the first and second experiment. Because the manager chose minimum support= 20 and minimum confident=20 leads to a less search time and the loss of a large number of data and few frequent for data. Fuzzy Cluster index technique (FCI) takes less time for searching .The Fuzzy Cluster Index technique (FCI) and Fuzzy Rule having the equal elapsed time approx.

Table 3: comparison between the three techniques

Items	Fuzzy Cluster Index (FCI) Time/ S	Fuzzy Rule without Cluster Index Time/ S	Apriori Algorithm Time/ S	Mini Supp	Mini Conf
393665	7.5840	10.2561	24.6466	20	20
207199	6.1275	8.1139	18.1953	20	10
144808	3.2327	7.5252	12.5262	20	0
41658	0.5683	6.3683	9.4308	10	0

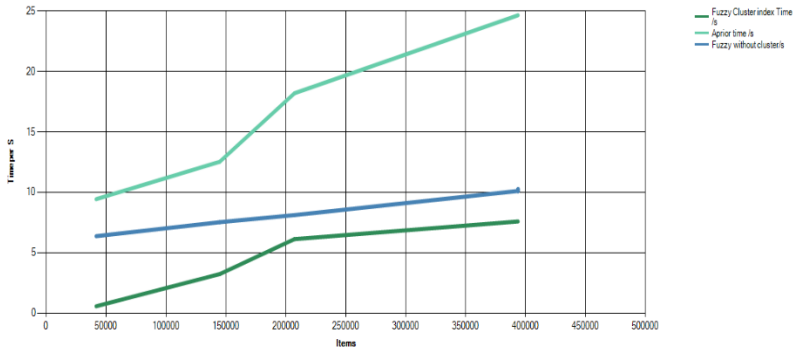


Figure 9: comparison between the three techniques

4.2.4 Fourth experiment:

In this fourth experiment, we show from Figure 10. Cluster index takes much time than other indexes and on the contrary, it takes less memory nearly 500 KB approx. Full text index consumes more memory and on the contrary, it gives best performance than others. After 5 lakh records in full text index, the memory size varies in less but after 8 lakh, memory size hike from 3000 KB to 5000 KB.[16].

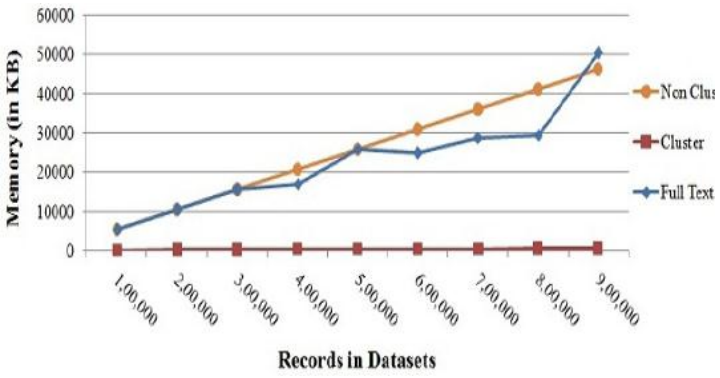


Figure 10: Indexes Memory consumption analysis [16]

5 Summary

In show [27] Figure 10 Cluster Index technique it consumed few memory comparatively other techniques which is costly for large data warehouse.

Note from the second and third experiment that the manager should the required expertise in data management because it determines the minimum support and minimum confidence.

Second experiment, the Index-based Apriori Algorithm time consumption is 24.6466 s. Because the manager chose minimum support= 20 and minimum confident=20 leads to a less search time and the loss of a large number of data and few frequent for data.

third experiment, the Index-based Apriori Algorithm time consumption is 232.7087 s, because the manager chose in Index-based Apriori Algorithm for minimum support= 2 and minimum confident=2 leads to a large search time and the loss of a small number of data and frequent large for data.

6 Conclusion and Future Work

Medical data warehouse is the central management system where, matching and searching are important operations. The traditional data warehouse was designed in such a manner that it can efficiently manage transactional data which is highly dominated by numerical information where as

in medical data warehouse textual and non-transactional information is encountered. The data set which contains text data is accessed over the network on the daily basis and performance issue arises. The aim of this paper is to propose Fuzzy Rule with an indexing technique based on few time for data warehouse used in application of medical field. The performance evaluation of three data warehouse queries is focused in this paper by comparing used with Fuzzy cluster index technique (FCI) and to observe the results of variable size data set with respect to time. Different three techniques in our proposed system has been used and analyzed using different types of queries on different size of data sets in medical data warehouse in order to perform operation in efficient manner. Fuzzy Cluster Index technique (FCI) provides better performance than Fuzzy Rule and Index-based Apriori Algorithm.

Future work can add fuzzy logic type 2 for improve performance data Warehouse by reducing the table or index fragmentation. And taking the views of users to identify possible queries that are used in the design of fuzzy rules.

REFERENCES

[1] Y. Bssil, “A Data Warehouse Design for A Typical University Information System”, LACSC – Lebanese Association for Computational Sciences Registered under No. 957, 2011, Beirut, Lebanon. Journal of Computer Science & Research (JCSCR) Vol. 1, No. 6, December 2012, pp. 12-17.

[2] R. Jindal1, and S. Taneja2 “COMPARATIVE STUDY OF DATA WAREHOUSE DESIGN APPROACHES: A SURVEY” 1 Associate Professor, Dept. of Computer Engineering, Delhi Technological University Formerly Delhi College of Engineering (DCE), Bawana Road, Delhi-42. 2 Research Scholar, Dept. of Computer Engineering, Delhi Technological University Formerly Delhi College of Engineering (DCE), Bawana Road, Delhi-42. International Journal of Database Management Systems (IJDMS) Vol.4, No.1, February 2012, pp.33-45.

[3] C. Gallo, and M. De Bonis, and M. Perilli “Data Warehouse Design and Management: Theory and Practice” IEEE MEMBERS. DIPARTIMENTO DI SCIENZE ECONOMICHE, MATEMATICHE E STATISTICHE UNIVERSIT`A DI FOGGIA Largo Papa Giovanni Paolo II, 1 - 71121 Foggia, Italy, Quaderno n. 07/2010, pp.1-18.

[4] D. Fasel, and K. Shahzad “A Data Warehouse Model for Integrating Fuzzy Concepts in Meta Table Structures” Information System Research Group Department of Informatics University of Fribourg Boulevard de Prolles 90 ,1700 Fribourg, Switzerland and Information Systems Laboratory Department of Computer and System Science Royal Institute of Technology (KTH) Forum 100, SE 164 40, Stockholm, Sweden,2010,pp.1-10.

[5] L. Sapir, and A. Shmilovici,”A Methodology for the Design of a Fuzzy Data Warehouse” Member IEEE, Lior Rokach, 2009, pp.1-8.

[6] M. Kokate, and Sh. Karwa, and S. Suman, and Prof. R. Chavan, "Performance Enhancement Techniques of Data Warehouse" International Conference on Advanced Computing and Communication Technologies, 2011, pp.67-72.

[7] A. Kumar, and D. Singh, and Dr. V. Sharma, "Achieving Query Optimization Using Sparsity Management in OLAP System" International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT),2014,pp.797-801.

[8] Dr. M. M. Hamad, and B. Anwer, "Knowledge-Driven Decision Support System Based on Knowledge Warehouse and Data Mining for Market Management" International Journal of Application or Innovation in Engineering & Management,Vol 3, 2014,pp.139-147.

[9] F. Kausar1, and Sh. Odah Al Beladi, and K. AL Shammari, "Comparative Analysis of Bitmap Indexing Techniques in Data Warehouse"International Journal of Emerging Technology and Advanced Engineering, VoL 4, no.1. 2014, pp.34-41.

[10] J. Zacek, and F. Hunka, “Data warehouse minimization with ELT fuzzy filter" Advances in Information Science and Applications, Vol II, 2014, pp.450-454.

[11] Igodan C. E, and Akinyokun O.C, and O. Olatubosun, "ONLINE FUZZY-LOGIC KNOWLEDGE WAREHOUSING AND MINING MODEL FOR THE DIAGNOSIS AND THERAPY OF HIV/AIDS" International Journal of Computational Science and Information Technology, Vol.1, No.3, 2013, pp.27-40.

[12] M. El-Wessimy, and H. M.O. Mokhtar, and O. Hegazy, "ENHANCEMENT TECHNIQUES FOR DATA WAREHOUSE STAGING AREA" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.6, 2013,pp.1-19.

[13] El. Aoulad Abdelouarit, and M. El Merouani, and A. Medouri, "The impact of indexes on data warehouse performance" International Journal of Computer Science Issues, Vol. 10, No 2, 2013, pp.34-37.

[14] K. Babar, and A. Gosain, "PREDICTING THE QUALITY OF OBJECT-ORIENTEDMULTIDIMENSIONAL (OOMD) MODEL OF DATA WAREHOUSE USING FUZZY LOGIC TECHNIQUE" [IJESAT], Vol 2, pp.1048-1054.

[15] Ms. R. Raval, and Prof. I. Rajput, and Prof. V. Gupta, " Survey on several improved Apriori algorithms" Journal of Computer Engineering (IOSR-JCE), Vol 9,2013,pp.57-61.

[16] A. Bansal, and S. Arora, "Performance Measurement of Indexing Techniques Used in Biomedical Databases" International Journal of Scientific & Engineering Research, Vol 3, 2012, pp.1-5.

[17] C. CIOLOCA, and M. GEORGESCU, "Increasing Database Performance using Indexes" Database Systems Journal, vol. II, no. 2, 2011, pp.13-22.

TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

Service Delivery Mechanism on Content Based Cluster Using Similarity of Services

[1]T.N.Anitha and [2]Balakrishna.R
[1]Department of CSE, S.J.C. Institute of Technology, Chickballapur, India
[2]Department of ISE, Rajarajeswari College of engineering, Mysore road, Bangalore, India
anithareddytn72@gmail.com; rayankibala@yahoo.com

ABSTRACT

Load balancing on web servers has become a major area of research due to ever increasing internet users' population and heavy load on popular website servers. Content based load balancing is proved to be a good mechanism to balance load on servers providing high quality services to the users' requesting for different category of content. On-demand creation of virtual servers has solved the complexity of load distribution and clusterization helps in grouping of same category servers. We propose a novel mechanism which works on clusterization of different grade servers intended to provide content based services to the users. Through experimental results it is found that this technique is helpful in increasing throughput and provides better quality of service to the users.

Keywords: Content based load balancing, Clusterization, ADC, Virtualization

1 Introduction

The flow of data traffic on internet is growing geometrically every year and so the load on cloud server to handle user requests. Yet users of internet expect page loading time to decrease due to the availability of high configuration systems at their end. It is very critical for popular websites to maintain high resource servers and new mechanism to keep response time to the lowest. Techniques such as virtualisation of servers, clusterization, load balancing, etc. are being studied to improvise the QoS (Quality of Service). A mix of clusterization and content based load balancing mechanism could be proved as a boon to provide high quality service to the website users.

A computer cluster is a set of connected systems, functioning in concert intimately in order that in several respects they create a single computer. The elements of a cluster are usually, but not at all times, linked to each other via rapid local area networks. Clusters are generally installed to enhance recital and/or accessibility over furnished by a single PC, whereas characteristically being very lucrative compared to single PCs of analogous speed or accessibility. Cluster Heads has the responsibility to make any interaction between its cluster members and ADC or server.

The computer comprises N uniform shared servers which offer the same documents, and a Cluster Domain Name Server (CDNS) which converts the URL-name into the IP-address of one of the servers in the cluster.

Clusters offer redundancy and distribution that make sure that website not at all goes down or loses vital operations or information. Clustered configurations permit simple scalability for parallel development, and are able to easily get a server offline for maintenance exclusive of compromise service. Particularly developed for businesses which insist high accessibility, clustered servers

DOI: 10.14738/tnc.32.1089
Publication Date: 7th April, 2015
URL: http://dx.doi.org/10.14738/tnc.32.1089

perform in recital for e-commerce websites, data-storage systems, internal networks, file and video distribution, high-volume blogs also other computing requirements. (Bader, David;, 1996)

1.1 Benefits of Dedicated Server Clustering

1.1.1 Redundancy and Trustworthiness

Diverse configurations in clustering could provide active or passive aspects in case one server breakdowns. Inactive choice comprises executing apps on a master committed server and containing a superfluous committed server to presume responsibilities if the master server breakdowns. In active configurations, two-server sets execute normal apps and represent from a general database with the intention that each server could occupy the other's responsibilities in cases of system breakdown.

1.1.2 Load Balancing

Configuring servers for utmost rapidity and recital when we have several traffic may needs dividing traffic and operations between servers for most favourable implementation. Targeted operations can be db, apps, storage systems or Web servers. Clustering permits us to perk up services radically, scale functions up or down hurriedly and identify cyber hacks prior to the reason for downtime.

1.1.3 High Accessibility

Clustering decreases singles points of breakdown and system susceptibility. Executing double load balancers, DBs, Web servers and superfluous network infrastructure avert downtime from break downs, cyber hacks, maintenance or natural calamities.

1.1.4 Data Growth

Irrespective of whether we manage a business, blog or aid or manage a data resource, a solitary committed server rapidly outgrows its processing and storage competencies. Having a clustering choice in place makes it simple to spread out without experiencing downtime which can reason for permanent harm to the status or loss of business.

1.1.5 Simple Maintenance

Server clusters permit for simple maintenance of the servers. If there is a trouble with a server, it could be detached from the cluster via either detaching the network wire or shout down the power. Once detached, the server could be repaired or reinstated. For the time being, the other servers in the cluster persist to execute processes providing as a minimum one server from the cluster relics online.

1.1.6 Rolling Upgrades

Server clusters make it simpler to upgrade servers or fix patches. As with any other maintenance, the servers in the cluster would go on with the essential processes, though merely one server from the cluster relics. Upgrading doesn't need downtime with a server cluster system.

2 Related Works

A generally hard problem in a shared setting is the recital squalor brought by an elevated load inequity and attaining lowest reply time for the clientele requests. Load balancing is hence vital for an assorted cluster, to promise a fair sharing of workload on every server in the cluster [1]. There are diverse methods for adopting load balancing in a shared assorted server setting. The taxonomy in [2] categorises the load-balancing methods into 4 groups: client-oriented, DNS-oriented, dispatcher-

oriented, and server-oriented methods. Every one of these methods largely executes load distribution algorithms that could be stagnant or dynamic and could utilize either centralized or shared control [3, 4, 5]. The Reference [6] represents that a hybrid of stagnant and dynamic approach for server choice offers a better recital. A client-oriented approach adopts the server choice on the clientele side [7]. The clientele could opt one of the servers in random but this random choice strategy could not promise load balancing and server accessibility. Alternatively the destination instigated approach needs a server to seek clientele requests [8] (from the overloaded servers). In a DNS oriented method, DNS server turns into a restricted access and confines throughput limiting performance [9]. A dispatcher oriented method acts address mapping at address point. A dispatcher oriented method might adopt either packet rewriting [10] wherein case the transparency of address rewriting is acquired [11] or the HTTP redirection that initiates high transparency compared to network load balancing, directing to weakening in performance.

2.1 History of Clusters

Greg Pfister has declared that clusters weren't discovered by any particular purveyor but by clientele who couldn't keep all their work on single system, or required a backup.[12] Pfister projects the date as some time in the 1960s. The official engineering base of cluster computing as a way of performing analogous exertion of any kind was debatably discovered by Gene Amdahl of IBM, who in 1967 printed what has approached to be viewed as the seminal paper on parallel processing: Amdahl's Law. (Bader, David;, 1996)

The history of near the beginning computer clusters is relatively directly attached into the history of early networks, as one of the main inspirations for the improvement of a network was to connect computing resources, forming a de facto computer cluster.

The foremost business clustering product was Datapoint Corporation's "Attached Resource Computer" (ARC) system, designed in 1977, and utilizing ARCnet as the cluster interface. Clustering as such didn't actually impression until Digital Equipment Corporation introduced their VAXcluster product in 1984 for the VAX/VMS OS (at present called as OpenVMS). The ARC and VAXcluster products not merely supported parallel computing, other than also distributed file systems and tangential tools. The thought was to give the benefits of parallel processing, whereas maintaining data dependability and exclusivity. Two other remarkable before time business clusters were the Tandem Himalayan (a circa 1994 high-ease of use product) as well as the IBM S/390 Parallel Sysplex (and circa 1994, mainly for commercial purpose). (Erguvan at el, 2009)

Within the same time framework, while computer clusters employed parallelism outer the computer on a product network, supercomputers started to utilize them within the same PC. Following the victory of the CDC 6600 in 1964, the Cray 1 was released in 1976, and launched interior parallelism by means of vector processing (Sedayao at el, 2008). Whereas in the early hours supercomputers expelled clusters and dependent on distributed memory, after a while some of the best ever supercomputers (such as K computer) dependent on cluster architectures.

2.2 Challenges in Cluster Computing

Load balancing consists getting the least loaded and paramount appropriate device in the network to run a work. In a local cluster setting this could simply be attained by centralized match making algorithms for example the one adopted in Condor (Erguvan at el, 2009). Centralized universal load creation pooling and drawing load-balancing decisions would not be realistic and measure well in universal cluster setting.

Dividing the universal cluster setting with a little interrelate topologies is the major for forming flexible cluster computing methods. Well-harmonized shared load balancing algorithms and protocols throughout clusters is essential to create most favorable resource distribution and utilization in an overall cluster setting. Even with such a system, getting the preeminent machine to carry out all jobs in global scale is not advantageous. One should observe the swapping among local and global optimization taking into consideration the price of stirring a job to distant sites for implementation. Meager cross-cluster implementation decisions may cause network overcrowding, and making systems inactive for long hours whereas making global appointment decisions and relocation of jobs. (Franco Milicchio, Wolfgang Alexander Gehrke, 2007)

Job scheduling comprises getting most appropriate profession to execute from amongst jobs belonging to numerous users and groups. In a local cluster setting this is accomplished by fair-share and main concern on the basis of setting up plans. Expanding this idea to divisions and projects extend throughout numerous geographic areas is the major for worldwide job scheduling optimization. Utilizing these global scheduling plans one should be competent to manage and implement project resource main concerns with challenging projects. This is needed for "good" throughput in global level. We refer better throughput since good consumption of resources for vital assignments as described by the user community. (Erguvan at el, 2009)

The algorithms of load-balancing which select the implementation place must also think about network stack, accessibility of user information, and safe implementation setting with same user qualifications all over the sites. These issues could be resolved with shared file mechanisms and computing setting services for example AFS, DCE/DFS, and Kerberos validation methods. (Sedayao at el, 2008)

Sharing of resources could be made in the vicinity at every site and through a centralized group for the universal system. As every site shares its own resources, it might permit users from other sites to distribute its surplus resources but provides foremost preference to the native users. When making use of a variety of the two strategies it might be feasible to have some of the resources assured for native users whereas remaining resources distributed among all other sites. The system has to be competent to implement these resource distribution strategies as the global distribution and operation snapshot is accessible at all stages. (Sedayao at el, 2008)

2.3 Architecture

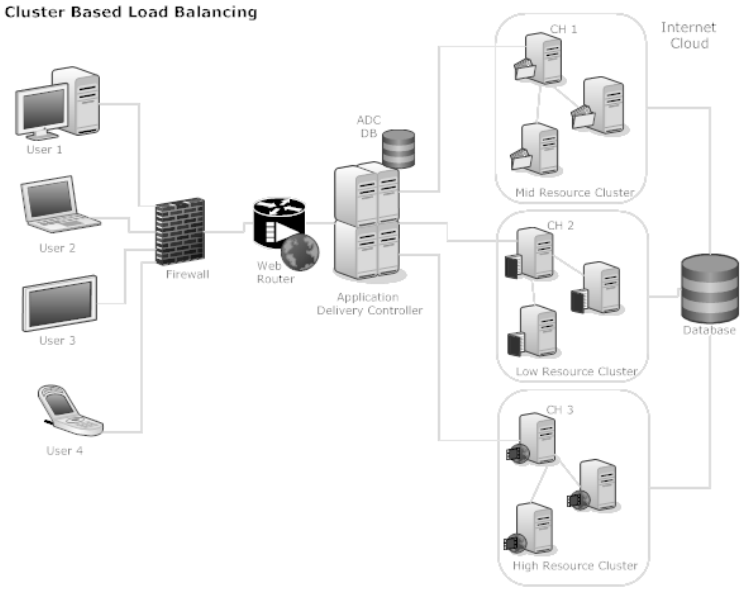


Figure 1: Cluster Based Load Balancing

3 Methodology

In the proposed mechanism, Multiple Virtual Servers (VS) are derived from Physical Servers (which is having capacity to serve .1 million user requests at an instance). VSs belonging to same set of configuration forms cluster. A cluster can keep maximum 382 VMs. User sends a URL request on browser to access a web application. After filtering through firewall it reaches to switch or router within LAN. URL Request is sent through ISP (Internet Service Provider) to the ADC Server. ADC analyses the requested content in User Request. ADC retrieves the list of Clusters belonging to that grade. ADC dispatches the request to Cluster Head (CH) of lowest load cluster. CH retrieves list of VMs within it and checks for lowest load VM. CH pings and checks availability of that VM, if available, then forwards User Request to the VM, else, searches for second lowest load VM which is available.

3.1 ADC

ADC or Application Delivery Controllers are high end load balancer devices used for distributing load among available virtual servers to enhance performance of any applications running these servers. These devices work on mechanism of receiving, analyzing and dispatching user requests to servers.

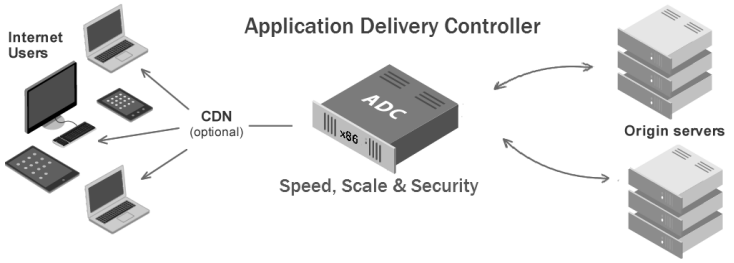


Figure 2: Application Delivery Controller

3.2 Cluster Formation

We propose a mechanism where clusters are formed dynamically as a result of excess amount of virtual servers in any cluster. We take a scenario where we assume a particular Physical Server (PS) which can handle at most .1 million User Request (UR) at a time.

$$Capacity(PS) \cong 100000\ UR$$

In our design, Clusters are formed out of same configuration VMs regardless of its distance and other parameter. The capacity of cluster to hold number of VMs is 382.

$$C_i = \{VM_1, VM_2, .., VM_{382}\}$$

Any excess of VM in a Cluster can lead to another cluster formation with exceeding number of VMs. Formation of Cluster is on-demand and it dynamically forms or dissolve whenever VM joins or leaves the group.

$$C_i\{VM_1, VM_2, .., VM_{382+1}\} = C_i\{VM_1, VM_2, .., VM_{382}\}, C_{i+1}\{VM_1\}$$

From each physical server multiple virtual servers are derived which resides in clusters for different configuration sets so that efficient services can be provided to all sort of user requests.

$$PS \ni \{C_1, C_2, .., C_n\}$$

Now as per the theorem we may say that total no. of requests served by all clusters belonging to one Physical Server could be equal to or less than Capacity of PS i.e. 0.1 million.

$$\sum UR \in (C_1 + C_2 + .. + C_n) <= (Capacity(PS) = .1m)$$

3.3 Service based Clusters

The configuration of Virtual Servers can be divided in 3 major categories i.e. high resource, medium resource and low resource VMs. Clusters are formed by joining similar configuration VMs and is completely on-demand. As no. of requests grow, new VMs are dynamically created and kept inside cluster. There could be 'n' no. of clusters as it depends upon which type of clusters are more. For e.g. clusters of low configuration set of VMs would take much less resource than clusters of high configuration VMs.

3.4 Algorithm Used

Table 1: Notations

Notation	Description
PS | VS	Physical Server | Virtual Server
G	Grade of Server as per Resources
CL	Cluster of Same Grade Servers
Req	URL Request
CT	Content Type in Request
ADC	Application Delivery Controllers
Req_k	Specific URL Request in ADC
SL | S	Servers List | Server
n	Upper Bound of Server List
LDS | DS	List of Down Servers | Down Server
A	Availability of Server
A_{s_i}	Availability of ith Server, where i>=1 and i<=n
L | L_{s_i}	Load | Load on Particular Server
L_{min}	Minimum Load
$L_{min_{s_i}}$	Server with Minimum Load

1. $PS = \{VS_1, VS_2, \ldots, VS_n\}$
2. $\{VS_1, VS_2, VS_n\} \in G_1, \{VS_1, VS_2, VS_n\} \in G_2, \{VS_1, VS_2, VS_n\} \in G_3$
3. $CL_1 = \{SL(G_1) \leq 382\}, CL_2 = \{SL(G_2) \leq 382\}, CL_3 = \{SL(G_3) \leq 382\}$

4. $DS \leftarrow \emptyset,\ L_{min} \leftarrow 1000$
5. $PS \leftarrow \sum Req \leq .1million$
6. $ADC = \{Req_1,\ Req_2, .., Req_m\}$
7. ***for all*** $Req \in ADC$ ***do***
8. $GetGrade(CT(Req)) \equiv G_x$
9. $Req \rightarrow minLoad\big(CL(G_x)\big).CH$
10. ***for all*** $S \in CL_x$ ***and*** $i \leq n$ ***do***
11. $L_{s_i} \leftarrow \sum Req\ \in S_i$
12. ***if*** $L_{s_i} \leq L_{min}$ ***then***
13. $L_{min} \leftarrow L_{s_i}$
14. $L_{min_{s_i}} \leftarrow s_i$
15. ***end if***
16. ***end for***
17. $A_{s_i} \leftarrow INetAddress.isAvailable\left(L_{min_{s_i}}\right)$
18. ***if*** $A_{s_i} \equiv true$ ***then***
19. $Req_k \rightarrow L_{min_{s_i}}$
20. ***end if***
21. ***else***
22. $DS \leftarrow L_{min_{s_i}}$
23. $LDS \leftarrow LDS \cup \{DS\}$
24. $LS \leftarrow LS - LDS$
25. ***Repeat*** $Step\ 10$
26. ***end else***
27. ***end for***

3.5 Random Walk Algorithm

Table 2: Random Walk Algorithm

Notation	Stands For
G	Graph
V	Vertices
E	Edges
VS	Virtual Server
t	Time
P	Probability
μ	Allocation of Server

Random walk algorithm works on probability distribution of virtual servers to be visited. It can be integrated in various kinds of P2P and web applications where multiple servers exist and each server need to be visited randomly. We propose this mechanism to choose a Cluster or Virtual Server from it to serve User Requests. Let us assume to have a Graph which is a set of Vertices and Edges which are interconnected. So, it can equated as

$$G = (V, E).$$

Beginning of the random walk happens with virtual server VS_0 that is either already defined to be picked first or kept in that position at the time of setup. Now as the random walk goes further to

Virtual Sever VS_1 at time t, it reaches to neighbouring Server VS_{t+1} at time $t+1$ which is selected at random with a definite probability allocation μ. Assume μ_t represent the allocation of virtual serverVS_t, so that

$$\mu_t(m) = P(VS_t = m)\ \forall\ m \in V.$$

Let

$$P = \left(P_{m,n}\right), m, n \in V$$

Depict the Random Walk's evolution matrix, where $P_{m,n}$ is the possibility that random walk switches from virtual server m to Virtual server n in single go. $P_{m,n} = 0$ if virtual servers m, n are not neighbours. The modus operandi of random walk is

$$\mu_{t+1} = \mu_t P = \mu_0 P^{t+1}.$$

4 Experimental Results

Various results were obtained duing experiments condected using JMeter Software on hetrogeneous systems. In order to have 2 levels of request dispatching, one system was made to act as ADC which was responsible for analyzing content requested and finding minimum load cluster of that grade and forwarding request to it's cluster head. Secound level of dispatching happens through cluster head which gets status of all its member servers and after checking availability forards request to it. To analyze effect of load balancing different amount of request loads were produced on servers using JMeter.

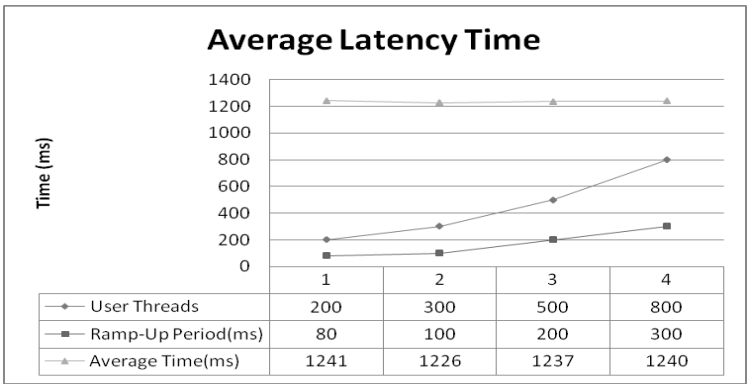


Figure 3: Average Latency Time

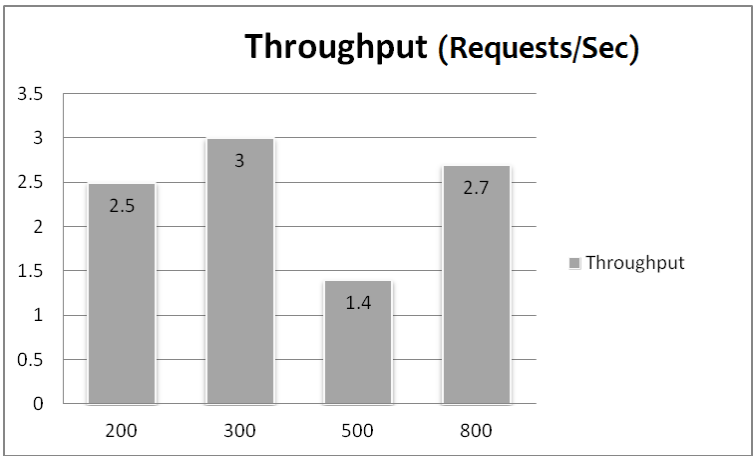


Figure 4: Throughput (Requests/Sec)

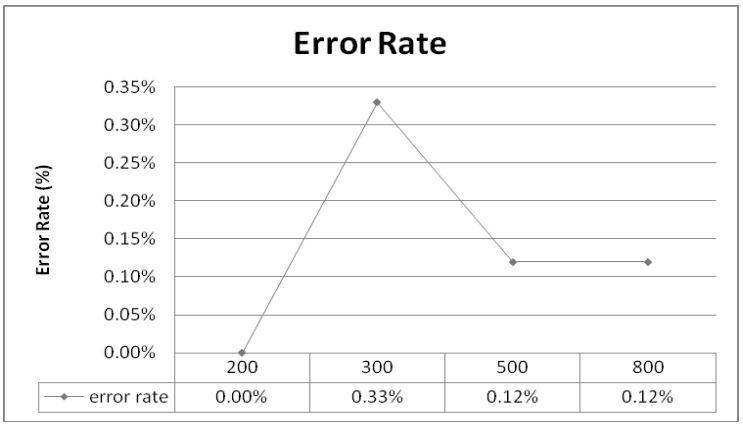


Figure 5: Error Rate

Fig 3 states the latency in serving use requests. It shows latency of 1241 milliseconds in handling each user request when load amount to 200 user requests, 1226ms for 300 user requests, 1237ms for 500 user requests and 1240ms for 800 user requests. It shows there is no particular effect of load due to increasing no. of user requests as the processing time remains approximately similar. Figure 4 represents throughput per second which is quite tend to vary with the processing speed of systems over time. It shows handling of average of 2.5, 3, 1.4 and 2.7 requests per second with error rate (ref. Figure 5) of 0%, .33%, .12% and .12% for 200, 300, 500 and 800 user requests respectively.

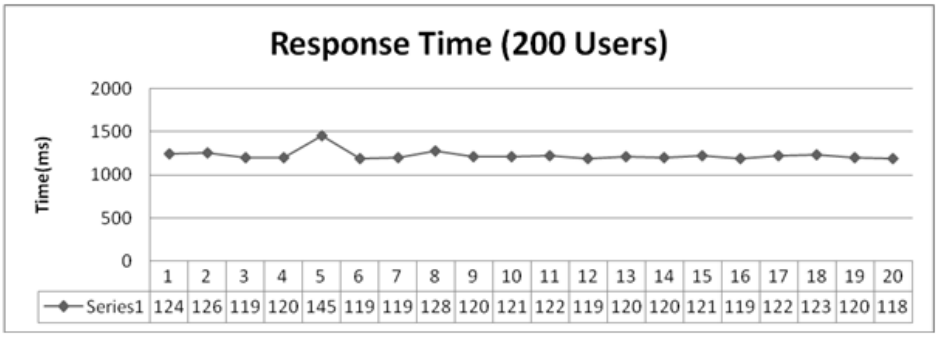


Figure 6: Response Time (200 Users)

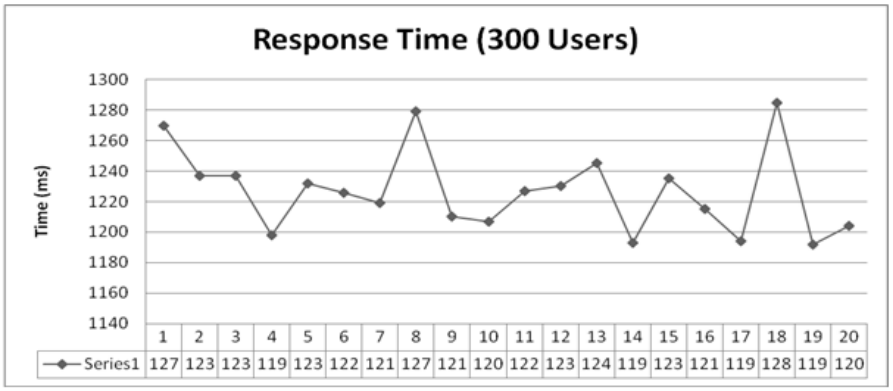


Figure 7: Response Time (300 Users)

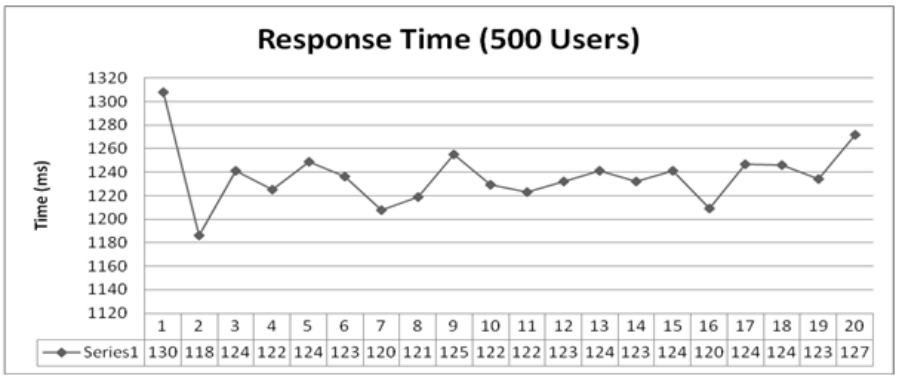


Figure 8: Response Time (500 Users)

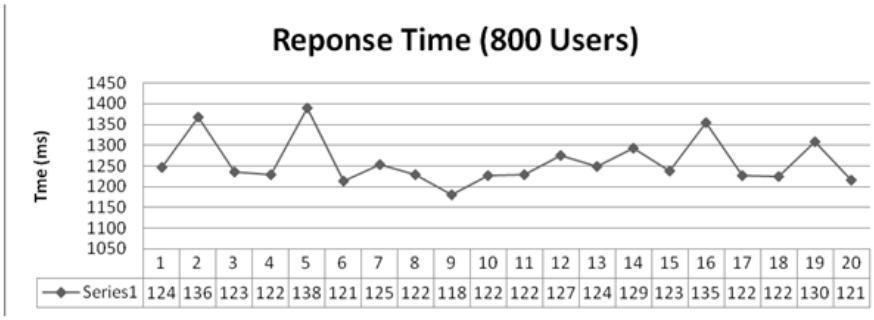


Figure 9: Response Rate (800 Users)

5 Conclusion

Content based load balancing has achieved immense weightage in the area of reasearch on load balancing on web and cloud servers. Clusterization techniques were found to be very much helpful in organizing virtual servers as per the ranking alloted based on their resource configurations. Experimental results on hetrogeneous systems has indicated this mechanism is helpful to achieve a good throughput with less error rate while doing load balancing on clustered servers. As there are 2 levels of dispatching required so initial server allocation takes little extra time but as this difference comes in milliseconds and can be ignored as it acts as preliminary to provide better QoS and satisfactory usage of service as per the content request.

Future works can be carried out to make this system more efficient and without any error rates by using additional improved mechanisms.

REFERENCES

[1] W. Tang, and M. Mutka, "Load distribution via static scheduling and client redirection for replicated web servers," Proceedings of the First International Workshop on Scalable Web Services (in conjunction ICPP 2000), Toronto, Canada, pp. 127-133, 2000.

[2] V. Cardellini, and M. Colajanni, "Dynamic load balancing on web-server systems," IEEE Internet Compute 3, pp. 28-39, 1999.

[3] T.L. Casavant, and J.G. Kuhl, "A taxonomy of scheduling in general-purpose distributed computer systems," IEEE Trans. Software Eng. 14 (2), pp.141-153, 1988.

[4] J. Cao, G. Bennett, and K. Zhang, "Direct execution simulation of load balancing algorithms with real workload distribution," J. System. Software 54, pp. 227-237, 2000.

[5] Y. Wang, and R. Morris, "Load sharing in distributed systems," IEEE Trans. Compute. C-34 (3), pp. 204-217, 1985.

[6] M.J. Zaki, W. Li, and S. Parthasarthy. "Customized Dynamic Load Balancing for a Network of Workstations," Journal of Parallel and Distributed Computing 43, pp. 156-162, 1997.

[7] C. Yoshikawa, B. Chun, P. Eastham, A. Vahdat, and T. Anderson, "Using smart clients to build scalable services," Proceedings of USENIX, pp. 105-117, 1997.

[8] D. Eager, E. Lazowska, and J. Zahorjan, "A comparison of receiverinitiated and sender-initiated dynamic load sharing," in Perform. Eval.Vol. 1, pp. 53-68, 1986.

[9] Cisco Distributed Director, http://www.cisco.com/warp/public/cc/pd/cxsr/dd/index.shtml.

[10] A Bestarvros,M Crovella, JLiu,and D. Martin, "Distributed packet rewritting and its applications to scalable web server architectures," Proceeding of the Sixth International Conference on Network Protocols, Austin, TX, pp. 290-297, 1998.

[11] D. Dias, W. Kish, R. Mukherjee, and R. Tewari, "A scalable and highly available web-server," Proceedings of the 41st International Computer Conference (COMPCON'96), IEEE Computer Society, San Jose, CA, pp. 85-92, 1996.

TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

Segmentation of Broken and Isolated characters in Handwritten Gurumukhi Word using Neighboring pixel technique

Akashdeep Kaur, Paramjeet Singh and Shaveta Rani
Giani Zail Singh Punjab Technical University Campus, India
Akashbrar702@yahoo.com

ABSTRACT

Character Segmentation of Handwritten Documents has been an active area of research and due to its diverse applicable environment; it continues to be a challenging research topic. In this paper, the focus is on offline segmentation of handwritten documents written in Gurumukhi Script. The desire to edit scanned text document forces the researchers to think about the optical character recognition (OCR). OCR is the process of recognizing a segmented part of the scanned image as a character. OCR process consists of three major sub processes - pre processing, segmentation and then recognition. Out of these three, the segmentation process is the most important phase of the overall OCR process. In this paper, algorithm is formulated to segment the scanned document image as a character that can be isolated or broken from within the given word. According to proposed algorithm, one part is extracting line from a document other part is extracting a word from the line. Segmentation part of the algorithm extracts characters from the extracted word. To segment the characters from a word, combination of two approaches which are Horizontal Profile Project and Vertical Profile Projection is used and will formulate a new algorithm which is Neighboring Pixel algorithm for touching characters in a word written in Gurumukhi script.

Keywords— Segmentation, Feature Extraction, Binarization, Classification, proposed work, Results

1 Introduction

Transmission and storage of information is done not only through computers but also through paper documents. To integrate these two mediums of information flow, a solution is for computer to "read" paper documents. Machine simulation of human reading is one of the areas, which has been the subject of intensive research for the last three decades, yet it is still far from the final frontier. So, works are still going on this direction.

1.1 Natural Language Processing

Natural language processing is a field of science and linguistics concerned with the interaction between the Computers and human languages. Natural language generation systems convert information from computer databases into readable human language. The term "natural" language refers to the languages that people speak, like English and Japanese and Hindi, as opposed to artificial languages like programming languages or logic. "Natural Language processing", programs that deal with natural language in some way or another. The study of human languages developed the concept of communicating with non-human devices.

DOI: 10.14738/tnc.32.1094
Publication Date: 15th April, 2015
URL: http://dx.doi.org/10.14738/tnc.32.1094

NLP deals with the Artificial Intelligence under the main discipline of Computer Science. The goal of NLP is to design and build software that will analyze, understand and generate languages that humans use naturally.

There are many applications of Natural Language processing developed over the years. The main applications are text-based, which involves searching for a certain topic or a keyword in a large document, translating one language to another or summarizing text for different purposes.

2 Character Segmentation

Character segmentation is the term, which covers all types of machine recognition of characters in various application domains. The intensive research effort on the field of character segmentation was not only because of its challenge on simulation of human reading, but also, because it provides efficient applications such as the automatic processing of bulk amount of papers, transferring data into machines and web interface to paper documents . A character segmentation system can be either “online” or “offline.” According to the mode of data acquisition, character segmentation methodologies are categorized into two systems as:

Online character segmentation systems

Offline character segmentation systems

2.1 Online character segmentation systems

Online character segmentation is the process of segmenting handwriting, recorded with a digitizer, as a time sequence of pen coordinates. It captures the temporal and dynamic information of the pen trajectory. Applications of on-line character segmentation systems include small handheld devices, which call for a pen-only computer interfaces and complex multimedia systems, which use multiple input modalities including scanned documents, speech, keyboard and electronic pen. These systems are useful in social environments where speech does not provide enough privacy. Pen based computers, educational software for teaching handwriting and signature verifiers are the examples of popular tools utilizing the on-line character segmentation techniques.

2.2 Offline character segmentation systems

Offline character segmentation is the process of converting the image of writing into bit pattern by an optically digitizing device such as optical scanner or camera. The segmentation is done on this bit pattern data for machine-printed or handwritten text. Applications of offline segmentation are large-scale data processing such as postal address reading; check sorting, office automation for text entry, automatic inspection and identification. Offline character segmentation is a very important tool for creation of the electronic libraries. Also, the wide spread use of web necessitates the utilization of offline segmentation systems for content based Internet access to paper documents.

3 Binarization

3.1 Scanning image:

In this step the document is converted into scanned image with the help of image scanner.

3.2 Binarization:

In this step gray scale images are converted to binary image with the help of OCR Software [10]. The images that are scanned are in the grey tone. Basically a Binarization is the process in which the grey scale images are converted into binary form means in the form of 0’s and 1’s.

Binarization separates the foreground (text) and background. There are various methods for binarization but the most common method for binarization is to select the proper threshold for the intensity for an image and then convert all the intensity values above the threshold to one intensity value (white) and all intensity values below the threshold to other chosen intensity (black).

4 Literature Survey

Vikas J Dongre, Vijay H Mankar[7] in 2010,"A Review of Research on Devnagari Character Recognition", in this paper, recognition of handwritten character is presented. There are five steps for the recognition of character recognition: 1) Pre-processing of image 2) Segmentation of words into characters 3) Feature Extraction 4) Reorganization 5) Post- processing.

Naresh Kumar Garg, Lakhwinder Kaur & M.K. Jindal [8] in 2011,"The Hazards in Segmentation of Handwritten Hindi Text" ,OCR is used to recognize the scanned text that can be in the form of handwritten or typed form. Segmentation is the important phase in the character recognition that can improve/decrease the accuracy of character recognition. Segmentation of printed words is quite easy as compare to handwritten words because of the various problems that will occur in the segmentation of handwritten text. There are two types of problems that can occur in the segmentation of handwritten text: 1) The Problems that can be ignored (Like the problems due to speed of writing). 2) The Problems that cannot be ignored.

Ashwin S Ramteke, Milind E Rane [9] in 2012,"Offline Handwritten Devanagari Script Segmentation", the process of Segmentation is a vital phase in the recognition of text. Devanagari is very useful Script in India. The segmentation of devanagari words is very difficult due to the presence of large character set that include consonants, vowels and modifiers. In this paper the major focus was on the segmentation of line, word and characters. Before the segmentation of an image some pre-processing of the image is done using the median filter and it also includes the binarization and scaling of image. After this pre-processing the segmentation is done. For the Segmentation of handwritten Devanagari script the histogram of input image is generated that shows the space b/w the characters so from this the characters can be segmented.

5 Identify the Presence of Broken Characters

Now after the segmentation of various characters the next step is to find that whether there is any broken character or not. Character can be broken due to writer's pen or page quality used. Segmentation of the broken character is quite difficult because vertical profile projection technique assumes the broken parts of the characters as individual characters and thus segmenting the word as separate character. So neighboring pixel technique is used to identify the broken character. So by concentrating on this feature, following steps are performed:

- Check the Neighboring pixels on both left and right side.
- If the black pixels are there then that represents the character is broken and not to be segmented.
- But if there are white pixels in its neighbor then these
- Pixels are treated as a gap and hence to be segmented.

Figure 1: Identification of broken character

6 Segmentation of Broken Characters

Now after the previous steps it is determined that which character is the broken character. So now there is need to make that broken characters as one character. This is done by scanning the neighboring pixels before segmenting the word into character. For this, following steps are performed:

- For each ith column of the word
- If all the pixels are white and if so then check i-1 and i+1 number of pixels.
- If all three pixels are white then treat them as gap between two characters and then segment the word.
- Check for the two pixels (i-1, i+1)
- If they are black, than it represents the broken character and don't segment the word from the ith pixel.

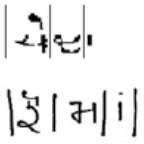
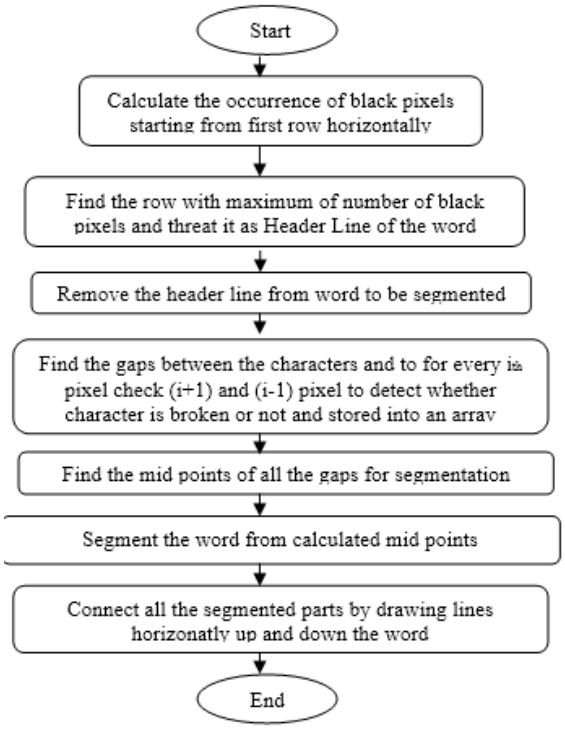


Figure 2: Segmentation of broken words

Neighboring pixel algorithm that can segment isolated, broken character is shown as below:



7 Results

In order to detect and segment broken characters in scanned word of handwritten Gurumukhi script documents, neighboring pixel have been used. This technique has been applied on the documents of three different categories. The category wise results of segmentation accuracy are given in table.

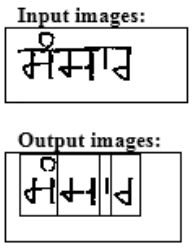


Figure 3: Input and Output images of broken words

8 Disscussion and Conclusion

Here this algorithm has been tested on 25 handwritten words taken from different people with different handwriting. In which there was isolated and broken characters.

Table 1: Different phases of words showing accuracy.

Phases	Words	Correctly segmented	%age
Phase 1:Words without an broken characters. (ISOLATED)	**75**	**75**	**100%**
Phase 2:Words with isolated, broken in one word(BROKEN)	**50**	**46**	**94%**

In the second phase the words with broken characters are handled and of these 46 (94% of 50) words are properly segmented and the remaining (6%) error was primarily because of overlapping characters with broken characters. The errors of over-segmentation were unavoidable because of the gaps in the broken characters. Any readjustment of the threshold value leads to high degree of under-segmentation in the words and therefore is not recommended.

REFRENCES

[1] G.S lehal, and Chandan singh, "A post-processor for Gurmukhi OCR", in Sadhana Vol. 27, Part 1, February 2002, pp. 99–111. © Printed in India

[2] G.S lehal and Daramveer sharma, "An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurmukhi Script"The 18th International Conference on Pattern Recognition (ICPR'06), IEEE 2006.

[3] G.S lehal, R. K. Sharma, and M. K. Jindal, "Segmentation Problems and Solutions in Printed Degraded Gurmukhi Script", World Academy of Science Engineering and Technology Volume 2, 2008.

[4] Naresh Kumar Garg, Lakhwinder Kaur and M.K. Jindal "Segmentation of Handwritten Hindi Text" International Journal of computer Applications, vol. 1-No. 4,pp19-23,2010.

[5] Galaxy Bansal, Daramveer Sharma,"Isolated handwritten words segmentation techniques in gurumukhi script" International Journal of Computer Applications (0975 - 8887), Volume 1 – No. 24, 2010.

[6] Vijay Kumar, Pankaj K. Sengar," Segmentation of Printed Text In Devanagari Script And Gurmukhi Script" International Journal Of Computer Applications (0975 – 8887) Volume 3 – No.8, pp24-29 June 2010.

[7] Vikas J Dongre, Vijay H Mankar," A Review of Research on Devnagari Character Recognition" International Journal of Computer Applications (0975 – 8887), Volume 12– No.2,pp8-15 November 2010

[8] Naresh Kumar Garg, Lakhwinder Kaur & M.K. Jindal ,"The Hazards in Segmentation of Handwritten Hindi Text" International Journal of Computer Applications (0975 – 8887) ,Volume 29– No.2, September 2011

[9] Ashwin S Ramteke, Milind E Rane,"Offline Handwritten Devanagari Script Segmentation"international Journal Of Scientific & Technology Research Volume 1, Issue 4,pp142-145, MAY 2012

[10] Gazal Munjal, Ms. Neha Sahu," Study of techniques used for Devanagri Handwritten Character Recognition" International Journal of Research in Engineering and Sciences(IJRES), Vol 1, Issue 2, pp.34-40, 2013.

[11] Simpel rani, Arbha Goyal ,"An efficient approach for segmentation of touching characters in handwritten hindi word"International conference oon Information and mathematical Sceinces 2013, 2014 ELESVIER.

[12] Munish kumar , Mk jindal , R.K.Sharma, "segmentation of Isolated And Touching Characters in Offline Handwritten Gurumukhi Script Recognition" I.J. Information Technology and Computer Science, 2014, 02,58-63.

TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

Investigating Privacy Preserving Healthcare Social Network

Qurban A Memon, Asma Fayes and Mustafa
College of Engineering, UAE University, Al-Ain, UAE
qurban.memon@uaeu.ac.ae

ABSTRACT

As health information is valuable, thieves will continue to steal it. Likewise, insufficiently trained employees at organizations or even individuals that pay less attention to creating a privacy-aware culture will suffer data losses when unprotected devices are lost, stolen or sniffed. In this work, sensors are exploited to collect and generate data to be processed, clustered and shared via locally developed application to improve social networking in context of privacy preserving healthcare. The solution is developed for Android operating system. It provides the user with a set of network related social services. In addition to (i) showing specific places (ii) sharing the user location; (iii) showing nearby friends; (iv) getting and sharing weather temperature, it clusters nearby friends; calculates and shares distance moved, calories burned and active time; calculates tracking and shares the user heart rate,. Data privacy model is presented using data session levels and user roles to ensure privacy within healthcare social network.

Index Terms—Social Networks, Privacy, Role based access, Healthcare Network

1 Introduction

Wireless sensor networks (WSNs) have become mature enough for widespread adoption. However, most of the related use cases are highly application-specific, and do not usually affect everyday life. WSNs could be made more appealing to end users by leveraging online social networks (OSNs). This includes developing novel application scenarios and interaction paradigms between WSNs and OSNs. Nowadays sensors and social networks can fruitfully interface, from sensors providing contextual information in context-aware and personalized social applications, to using social networks as "storage infrastructures" for sensor information. The integration of sensor networks with social networks leads to applications that can sense the context of a user in better ways and thus provide more personalized and detailed solutions. Social networks have gained popularity recently with the advent of sites such as MySpace, Friendster, Facebook, etc. These networks are a source of data as users populate their sites with personal information. To better understand how online social networks can be integrated with physical world, there is a need to understand services provided by current OSN's, which are (i) identity and authorization services, (ii) Application Programming Interfaces (APIs) to access and manipulate the social network graph, publish and receive updates and (iii) container facilities for hosting third party applications [1].

There are a couple of important drivers for integrating sensor and social networks. One driver for integrating sensors and social networks is to allow the actors in the social network to both publish their data and subscribe to each other's data either directly or indirectly after discovery of useful information from such data. The idea is that such collaborative sharing on a social network can increase real-time awareness of different users about each other. A second driver for integrating sensors and social networks is to better understand or measure the aggregate behavior of self-

DOI: 10.14738/tnc.32.1116
Publication Date: 17th April, 2015
URL: http://dx.doi.org/10.14738/tnc.32.1116

selected communities or the external environment in which these communities function. Examples may include understanding traffic conditions in a city, understanding environmental pollution levels, or measuring obesity trends [2].

Some examples of integration of social and sensor networks may be exemplified as, for example the Google Latitude application, which shares the collected mobile position data of the user. As a typical use, the proximity alerts may be triggered when two linked users are within geographical proximity of one another. As another example, the City Sense application collects sensor data extracted from fixed sensors, GPS-enabled cell phones and cabs in order to determine where the people are, and then carries this information to clients who subscribe to this information. A number of real-time tracking applications such as 'Automotive Tracking Application' determine the important points of congestion in the city by pooling GPS data from the vehicles in the city. Animal tracking uses tracking data collected with the use of radio-frequency identifiers [2]. The CenceMe application injects sensing presence into popular social networking applications such as Facebook, MySpace, and IM (Skype, Pidgin) allowing for new levels of "connection" and implicit communication between friends in social networks [3]. The Green GPS is a participatory sensing navigation service that maps fuel consumption on city streets, to allow drivers to find the most fuel-efficient routes for their vehicles between arbitrary end-points [4].The Microsoft SensorMap allows for a general framework where users can choose to publish any kind of sensor data. The SensorMap application enables users to index and cache data. The indexing and caching allows users to issue spatio-temporal queries on the shared data [2].

Sensors provide numerous research challenges from the perspective of analysis. Since the collected data typically contains sensitive personal data (e.g., location data), it is extremely important to use privacy-sensitive techniques to perform the analysis. Another challenge is that the volume of data collected can be very large. For example, in a mobile application, one may track the location information of millions of users simultaneously. The innovations in World Wide Web [5, 6] and the recent trends in data protection [7, 8] have increased attraction for use of online social networks. However, the related advancements in technology and tools are also complimented by corresponding privacy concerns. The important thing to note is the lack of awareness for potential risks involved when data is being shared online [9]. Specifically, the external entities can mine this data and use it for different purposes like spamming [10], discovering interaction pattern in the enterprise to offer and develop innovative services, identification of the important person in the network, detection of hidden clusters, identifying user sentiments for proactive strategies etc. [11].

In context of healthcare, currently fitness bands, electronic health records, health information exchanges, connected devices, tools, and sites containing medical data keeps growing. Attackers are also becoming more sophisticated. Cybercriminals are seeking more information than ever about their victims to sell. The concern is that richer personal identity data of individual users, consisting regional and geographic data, personal information and behavior are expected to be traded in the same manner that stolen credit cards are today. Thus, the task of securing health information is becoming more challenging for even the best-prepared organizations.

The purpose of this research is to improve social networking in area of healthcare by sharing specific information (for example accumulated fitness indicators) online with doctor and/or parents in a privacy aware manner.

The paper is structured as follows. In the next section, a model is proposed that describes what is needed to build such an application, and which social parameters need to be linked, and how the

platform addresses privacy. The section three discusses the model implementation. The development details are discussed in section four along with results. The comparative analysis is carried out in section five, followed by conclusions in section six.

2 Proposed Model

A convenient design gives rise to other challenges that need to be addressed in order to enable development of successful mobile applications which meet user needs. Before describing the model, it seems necessary to highlight challenges in the environment:

- Mobile limitations: Mobile phones have very good computational efficiency, but they offer limited programming and resource usage control.
- Energy limitations: Application developers on mobile phone platforms need to be aware of power consumption when developing an application that depends on using radio interfaces such as GPS.
- Privacy issues: The collected data such as location data contains sensitive and personal data. It is highly important to apply a privacy model to maintain data security.
- Data management: The data collected is huge, so it is extremely important to efficiently process the huge amounts of data.
- Simplicity: The application should be easy to understand and thus target non-expert.
- Determining user location: GPS only works outdoors, and it quickly consumes phone battery power. Android's Network Location Provider helps to determine user location using Wi-Fi signals and cell tower. This strategy helps to provide the location details indoors and outdoors and uses less battery power. For effective use of battery, the application needs to deploy both strategies.

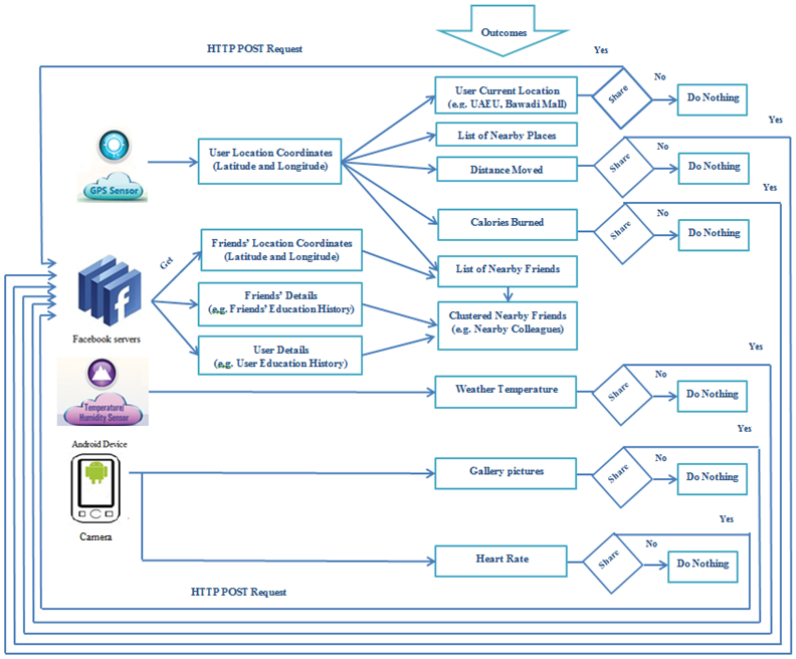


Figure 1: Conceptual Model

In the proposed model, the objective is to extend online social networks by injecting more sensing features like hear rate monitoring, sugar levels etc., and then processing and clustering some of this information for a benefit to a user in an environment like healthcare, using a locally developed application in the mobile device. The proposed model describes an application that enables members of a typical social network platform to share healthcare related features along with their

location information with their friends in a private manner. The application is expected to allow new levels of "connection" and implicit communication between contact groups in social networks. The application uses various sensors to acquire relevant data and display it on user device. The user can check-in to one of the displayed places, and this in turn, is named as a single visit to a location. Thus, the user can get his/her friends location based on their last check-in. The user should also be able to see if any of his/her friends are checked in nearby. For this, the displayed nearby friends need to be clustered to nearby colleagues, family and so on. In addition the application should enable the user to calculate the distance walked, the duration spent during the walk and the walking related burned calories. As a benefit, these services may motivate the user to exercise by competing with friends who exercised more and burned more calories. Moreover the user can check if any of his friends is nearby so he can walk with. The application is expected to encourage users to track their heart rates and share it with their doctor. As a result, this is expected to help the user to track their fitness to a new level. Based on this, a conceptual model is displayed as shown in Figure 1. Based on the model, the platform, components, and services that meet these expectations are detailed as follows:

2.1 Application Platform:

There are several popular OSN platforms. The Facebook is the most popular OSN platform today. This application is developed for such a platform and is compatible with Android devices. For development, the Facebook supports different APIs for developers: (i) The Graph API, which is a simple HTTP-based API that gives access to the Facebook social graph, uniformly representing objects in the graph and the connections between them. Most other APIs are based on the Graph API; (ii) The Open Graph API allows applications to tell stories through a structured, strongly typed API; (iii) The Facebook offers a number of dialogs for Facebook Login, posting to a person's timeline or sending requests; (iv) The Facebook Query Language (FQL) enables the developer to use a SQL-style interface to query the data exposed by the Graph API. It provides some advanced features not available in the Graph API such as using the results of one query in another; and (v) The Facebook Public Feed API lets the developer read the stream of public comments as they are posted to Facebook.

2.2 Application Sensors:

The application uses the built-in GPS of the user mobile device to get current location coordinates. It also uses the temperature and humidity sensor to check the weather temperature.

2.3 Application Services:

The services, which can be used using this platform, are:

2.3.1 Show nearby Places:

This application enables the user to use the built-in GPS or Androids' network location provider built in the mobile device to get the current location. It displays a list of nearby places as well.

2.3.2 Public Location Badge:

The user can post location directly on Facebook to increase visibility of information to other users.

2.3.3 Show nearby Contacts:

The user gets his contact location based on last Check-in. The user can see if any of his/her friends are checked in nearby. The application displays a list of nearby friends, place and the time they checked in.

2.3.4 Cluster nearby Contacts:

Clustering is important in analysis and exploration of data. This application clusters nearby friends into groups based on colleagues, family and so on.

2.3.5 Tracking Distance Moved, Calories Burned and Active Time:

This application tracks distance moved, calories burned and shows active time for the user. The application can also be used for running, cycling, walking and all other distance-based outdoor sports. Once data is shown on network, the user can seek extra encouragement from friends and family to workout. The clinical staff from a healthcare center, once connected in privacy mode, can also monitor shared clinical data.

2.3.6 Get weather temperature:

This application enables the user to use the built-in sensors of a mobile device to check the weather temperature and share with friends.

2.3.7 Share pictures:

The application enables posting image of user's specific injured or monitored body part, once on line, to be seen and examined by a doctor.

2.4 Application Security and Privacy:

As highlighted in section I, data privacy is a major concern. In order to maintain privacy, the user can turn off this whole or components of this application using the settings option, but this is not the common practice of online users. In order to protect users' shared data, data can be handled in many ways: (i) by bifurcating table into shareable and non-shareable columns i.e. data partitioning, and then play with rights. This will work only in the case when you exactly know which field to be shared and your table entries are fixed (i.e. not dynamic). However for data porting and scaling up the options, one may face problems; (ii) by using MongoDb. MongoDB is a no-SQL database and works on documents rather than entries. It's also hierarchical and supports dynamic data (i.e. table entries cannot be fixed); (iii) (Jugar option) by inserting one column of rights to all your data rows. In the first part of query, check is done to see whether rights are correct, then further processing of query is allowed, otherwise the query is rejected; and (iv) by using role based access control (RBAC) mechanism. In RBAC, permissions are associated with roles, and users are to be made members of appropriate roles [12-13]. This helps to simplify management of permissions. Roles are similar to the group's concept in access control. Role is defined as a set of users on one side and a set of permissions that will be applied to the users on the other side, while groups are defined as a set of users only. The privacy model made as part of the proposed scheme is derived from role based access control (RBAC).

In the proposed model, the roles are generated for various trust levels and contacts are assigned roles based on their relationship with the user. Contacts can be easily reassigned from one role to another. The developed application has predefined role-permission relationships, which makes it simple to assign contacts to the predefined roles. It is difficult, without the new privacy model, to determine what permissions have been assigned to what users. The proposed privacy model helps to perform large-scale authorization management. The basic concept of the proposed privacy model is that the user assigns his contacts to roles, roles have predefined permissions, and contacts acquire permissions by being members of roles. User-role can be many to many, which means that the same user can be assigned to many roles.

The application provides the user with three roles: (i) trusted, (ii) semi-trusted and (iii) un-trusted. The user will set the members of each role according to his/her relationships with his/her contacts. Trusted role has predefined permissions that will enable its members to view and comment on most of the user posts. Semi-trusted will enable its members to view and comment on some of the user posts, while untrusted role members won't be able to view or comment on many of the posts. Moreover, the user can assign relationships to his trusted and semi-trusted contacts. The user can assign his trusted contacts close or not close relationship. Assigning close relationship to contacts will give them the option to view more posts, while not-close relationship won't give them the permission to view more details. Using this privacy model, the users can allow their contacts to share certain level of their information. This model helps in managing different level of information sharing. Contacts will not know what role or relationship the user has assigned them. Such a model is shown in Figure 2.

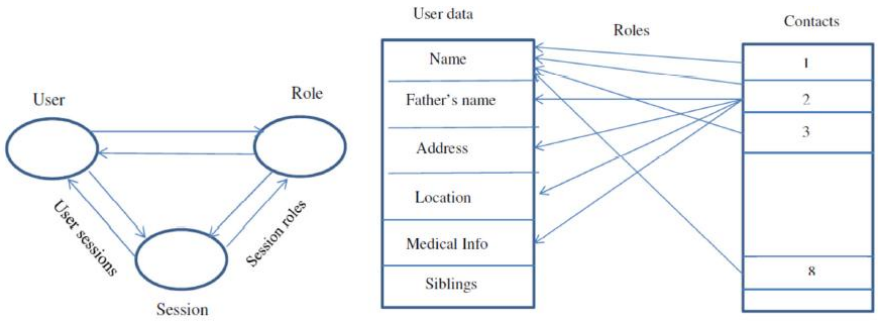


Figure 2: User-Data Model

It is clear from Figure 2 that each of the contacts has some level of access to the user data. For example, the contacts 1, 3, and 8 (i.e., friend, colleague, office staff) may only access name of the user, while contact 2 (a healthcare unit staff or user-parent) may also access address, location and medical records though all contacts may be part of same social network. In order to examine different relationships between entities, consider set of users, roles, objects, operations, collaborative relationships, access levels, and conditions be represented by *U, R, Objs, Oprs, CR, AL, Cond* respectively. Thus, the assignment relations among elements of the privacy-aware model are:

2^R: the power set operations

1. Many to many mapping user-role assignment relation: $URA \subseteq UxR$
2. The set of permissions: $Perms = 2^{(OprsxObjs)}$. This can also be stated as: $Perms = \{(Objs, Oprs) | Objs \in Objs, Oprs \in Oprs\}$
3. The set of sharing and privacy-aware permissions: $PA\_perms = (Perms, CR, Cond, AL)$ This can also be stated as: $PA\_perms = \{(perms, cr, al, cond) | perms \in Perms, cr \in CR, al \in AL, and\ cond \in Cond\}$
4. Many to many mapping permission-role assignment relation: $PRA \subseteq PermsxR$ Many to many mapping sharing and privacy-aware permission-role assignment relation: $Prv - PRA \subseteq PA\_perms\ xR$

3 Model Implementation

Based on the model in section II, the information and process flow inside application can be easily visualized. Based on this, software architecture is shown in Figure 3. For the purpose of implementation, the various stages in information and process flow are discussed below.

3.1 Sensing:

Most Android-powered devices have built-in sensors that measure motion, orientation, and various environmental conditions. These sensors prove raw data with precision and accuracy. The platform supports three broad categories of sensors: (i) position sensors to measure the physical position of a device. This category includes orientation sensors and magnetometers; (ii) environmental sensors to measure various environmental parameters, such as ambient air temperature and pressure, illumination, and humidity. This category includes barometers, photometers, and thermometers; and (iii) motion sensors to measure acceleration and rotational forces along three axes. This category includes accelerometers, gravity sensors, gyroscopes, and rotational vector sensors. In this application, position and environmental sensors are used.

3.2 Data Acquisition:

Data acquisition is the process of gathering information in an automated fashion from analog and digital measurement sources such as sensors. Apart from position and environmental sensors to get position (i.e., latitude and longitude information) and weather information respectively, the application uses Facebook APIs such as Graph API and Facebook Query Language to extract the user education history, user work history, friend's last check-in coordinates, friends education history, friends work history and the nearby places.

3.3 Data Processing:

The application analyses and processes the extracted data to produce meaningful information. After getting the user location coordinates and his/her friends' location coordinates, the application measures the distance between the user and each one of his friends to produce list of nearby friends. The application compares the user education-history and other details with friends' education-history to cluster the nearby friends into groups such as colleagues, family and so on. Additionally, the application uses the acquired location coordinates taken frequently to measure the distance walked, the related duration and the calories burned.

3.4 Data Sharing:

The application enables the user to share location, the distance walked, burned calories, the weather temperature and pictures. The Graph API updating is done simply with an HTTP POST request to relevant endpoint with the updated parameters. To publish and share new data, the application uses POSTs HTTP requests to appropriate URLs.

3.5 Presentation:

After processing the data, the application displays a list of nearby places. The user will be able to check in to any of these places by clicking on one of the places. It also has nearby friends icon by pressing on this icon, the user will get a list of his/her nearby friends based on their last check-in. The application organizes nearby friends into groups of colleagues, work friends, family and others. Moreover, it displays the distance the user walked, total calories burned and the weather temperature.

Figure 3: Software Architecture of application

4 Development Results

Before results are presented, it seems necessary to highlight development environment, briefly described below:

Simulation Environment: The application is developed for Android operating system devices. Android is an open source operating system available to all developers with various expertise levels. Android is a Linux-based operating system created for touchscreen mobile phones and tablets. Android platform allows users to develop, install and use their applications. The application is primarily developed in Java programming language by using the Android software development kit (SDK). The SDK provides the developers with all the tools they need including software libraries, debugger, sample code, emulator, and tutorials. The integrated development environment (IDE) for developed application is Eclipse using the Android Development Tools (ADT) plugin. Facebook SDK for Android was used to integrate this application with Facebook's platform.

In order to start the simulation, prerequisites including Eclipse, Android SDK, Android Developer Tools (ADT) Plugin, and Facebook SDK were installed. Then, Facebook SDK was imported to Eclipse. Every Android app that is developed was signed, as it was required to register each application's key hash with Facebook as a security check for authenticity. After registering on Facebook Developer Site as a developer, application Facebook profile was created and details such as application name, category, and key hash for were entered. After creating application profile, application ID was generated and appeared in the profile. Then application ID was added to project files on Eclipse.

In the following, the main focus is how all the functions, components and services mentioned in section II and III are implemented. The code and a sample view of some of the functions/components are also shown for clarity.

Login: This has been implemented in an easy way for people to log in to the application. The application uses iOS, Android, JavaScript and Facebook SDKs to speed up the process and build login systems quickly. For secure authorization Facebook uses the OAuth2.0 open protocol for confirming a person's identity and giving them control over right of access to their information.

Permissions: The permissions enable developers to request access to information about someone using their application. It asks for the following permissions: offline access, publish stream, publish check-ins, photo upload, user status, user education history, user work history, friends' status, friends' education history and friends' work history. To gain access, the application requests the permissions transparently through the Login dialog. To maintain information security, almost all API calls at Facebook need to have an access token passed in the parameters of the request.

Show nearby Places: The following steps outline how to get user's current location, display a list of nearby places and check in to one of these places with the Facebook SDK for Android.

a. Set up the Place Picker Item: This step includes defining a BaseListElement class to represent an item in the list. This class contains member variables that define the user interface (UI) as well as methods that are sub-classed to implement the behavior around click events, storing and restoring state info, as well as notifying observers about data changes.

b. Show the Places Picker: The Facebook SDK provides a placePickerFragment class that displays a list of nearby places. This fragment is hosted in the PickerActivity class. This activity launches when the user clicks on a place in the list. The PlacePickerFragment is used if the incoming intent data matches a pre-defined place picker Uri. Before loading the data, the

PlacePickerFragment is configured to specify search criteria like radius, query and maximum results to return.

c. Display the Selected Place: In this step, the place will be displayed when the place picker activity is dismissed.

Public Location Badge: The following steps outline how to publish a story to share the user location with friends. A request will be published by using Request(Session session, String graphPath, Bundle parameters, HttpMethod httpMethod). GraphObject and OpenGraphAction interfaces are used to set up a Graph object representation of the POST parameters. Facebook SDK is used to publish the user location by performing the following steps:

a. Construct a new Request for currently active session that is an HTTP POST to the me/checkins Graph API path.
b. Set a GraphObject for the Request instance. The GraphObject represents location parameters, like the selected place ID, message and location coordinates.
c. For best practices, the user is asked for publish_actions write permission in context, when the app is about to publish the user location.

Show nearby Contacts: To show nearby friends, Facebook Query Language (FQL) is used. FQL enables to use a SQL-style interface to query the data exposed by the Graph API. Below, the steps are described that show nearby friends.

a. Issue a HTTP GET request to /fql?q=query where query is a JSON-encoded dictionary of queries. The following code uses FQL to get the friends details and location according to their last check-in:

```
Bundle params = new Bundle();
params.putString("method", "fql.query");
params.putString("query", "SELECT author_uid,timestamp,coords,checkin_id FROM checkin WHERE
author_uid IN (SELECT uid2 FROM friend WHERE uid1 = me())  ");
String response = Utility.mFacebook.request(params);
response = "{\"data\":" + response + "}";
```

b. Store the friends' details and locations from JSONObject into the following arrays latitude, longitude, author_uid_array, timestamp and checkin_id. After that, the distance between the user and each one of his/her friends is calculated and stored in distances array.

The following code uses the response for the FQL query from the previous step and extracts the friends' details and locations from the response, and then it stores them in arrays. After that the distance is calculated and stored in distances array.

```
JSONObject json = Util.parseJson( response );
JSONArray data = json.getJSONArray( "data" );
JSONObject coords;
Long author_uid=(long)0;
for ( int i = 0, size = data.length(); i < size; i++ ){
JSONObject friend = data.getJSONObject( i );
if(author_uid!=friend.getLong("author_uid"))
{
coords = data.getJSONObject( i ).getJSONObject("coords");
latitude[counter] = coords.getDouble( "latitude" );
longitude[counter] = coords.getDouble( "longitude" );
author_uid = friend.getLong("author_uid");
author_uid_array[counter]=friend.getLong("author_uid");
timestamp[counter] = friend.getString( "timestamp" );
checkin_id[counter]=friend.getLong("checkin_id");
loc.       distanceBetween       (loc.getLatitude(),       loc.getLongitude(),latitude[counter],
longitude[counter], results);
distances[counter]=results[0];
```

```
counter++;}}
```

c. Find out nearby friends by comparing distance between the user and each one of his/her friends with a predefined distance, then store nearby friends' name in an array.

The following code uses the distances array from the previous step to compare the distance between the user and each one of his/her friends with a predefined distance-threshold in order to determine nearby friends. Additionally, the following code uses FQL to get nearby friends academics history details to be used in clustering.

```
for ( int i = 0, size = distances.length; i < size &&distances[i]!=0 ; i++ ){
if( distances[i]<(float)11500)
{
neededIndex[counter3]=i;
Nearby_friend_id[counter3]= author_uid_array[i] ;
counter3++;
}}
for ( int i = 0, size = counter3; i < size ; i++ ){
if(i!= size-1)
{
query+= "uid="+Nearby_friend_id[i]+" or ";
}
else
{
query+= "uid="+Nearby_friend_id[i];
}}
Bundle params2 = new Bundle();
params2.putString("method", "fql.query");
params2.putString("query", "SELECT uid,name,education FROM user WHERE "+query);
String response2 = Utility.mFacebook.request(params2);
response2 = "{\"data\":" + response2 + "}";
JSONObject json2 = Util.parseJson( response2 );
data2 = json2.getJSONArray( "data" );
for ( int i = 0, size2 = data2.length(); i <size2; i++ ){
JSONObject friend2 = data2.getJSONObject( i );
Nearby_friend_Name[i]= friend2.getString("name");}
```

d. Display nearby contact names, their exact location and when they checked in. For example: Dr. Noor Checked in Al-Ain Hospital at 2014-02-10 T09:16

In the following code, when user presses nearby friends' button, a list of nearby friends is displayed with their location and when they checked in. The list is shown in Figure 4.

```
Button mGetNearbyFriends = (Button) findViewById(R.id.get_nearby_friends);
mGetNearbyFriends.setOnClickListener(new View.OnClickListener() {
publicvoid onClick(View v) {
try{
TextView friends_Locations = (TextView) findViewById(R.id.friends_Locations);
friends_Locations.setText("" );
String jsonUser=null;
for ( int i = 0, size = counter3; i < size ; i++ ){
jsonUser= Utility.mFacebook.request(""+checkin_id[neededIndex[i]]);
obj                                =                                Util.parseJson(jsonUser);
       placeName[i]=obj.optJSONObject("place").getString("name");
created_time[i]=obj.getString("created_time");
friends_Locations.append(Nearby_friend_Name[i]  +"  checked  in  "+placeName[i]  +"  at  "+
created_time[i]+"\n");} }
catch (MalformedURLException e) {
e.printStackTrace();}
catch (IOException e) {
e.printStackTrace();}
catch (FacebookError e) {
e.printStackTrace();}
catch (JSONException e) {
e.printStackTrace();}
}});
```

Figure 4: Nearby friends list

Cluster Nearby Contacts: The following steps show, how clustering can be used to group nearby friends.

Issue a HTTP GET request, shown below, to /fql?q=query to get user academics history:

Issue a HTTP GET request, shown below, to /fql?q=query to get nearby colleagues after placing the user education history in the FQL query and display the nearby colleagues for the user.

Tracking Distance Moved, Calories Burned and Active Time: The following steps show, how tracking distance moved, calories burned and active time is calculated.

The application tracks the user moved distance by capturing user location coordinates periodically, calculating the distance between each two locations and summing the distances from the time the user presses start button till the time the user presses stop button.

The application will request the user to enter his/her weight in kilograms in order to calculate the walking burned calories. It uses the equation as shown below to calculate rate of calories burned per pound of body weight [14].

Rate per Pound (Cal/lb-min) = $A+BV+CV^2+KDV^3$ where:

V=Walking Speed (mph) – Limited to a minimum of 1 mph and a maximum of 5 mph
A= 0.0195
B= - 0.00436
C= 0.00245
D= $[0.000801(W/154)^{0.425}]/W$
W=Weight (lbs)
K= 0 or 1 (0=Treadmill; 1=Outdoors)

The code, shown below, uses the above equation to calculate the walking burned calories. When the user presses stop button, the application displays the distance walked, the duration spent during the walk and the walking burned calories. The Figure 6 shows resulting display on the mobile device.

```
privatevoid getDistanceAndCalories() {
TextView DistanceMoved = (TextView) findViewById(R.id.distance_moved);
TextView Activetime = (TextView) findViewById(R.id.active_time);
TextView WalkingBurnedCalories = (TextView) findViewById(R.id.burned_calories);
EditText Weight = (EditText) findViewById(R.id.weight);
DistanceMoved.setText("Total Distance you walked : "+ TotalDistance );
activeTime= (counter-1)*30/60; //minutes
Activetime.setText("Active time : "+ activeTime + " minutes \n");
activeTimeInHours=activeTime/60;
totalDistanceInMiles=TotalDistance/(float)1609.344;
```

```
A=(float) 0.0195;
B= (float)-0.00436;
C=(float)0.00245;
K= 1;
weightInKgs=(float)Double.parseDouble(Weight.getText().toString());
weightInPounds=weightInKgs*(float)2.20462;
D= (float)((Math.pow(weightInPounds/145, 0.425)*0.000801)/weightInPounds);
V=(totalDistanceInMiles/activeTimeInHours); //Walking Speed (mph) – Limited to a minimum of
1 mph and a maximum of 5 mph
ratePerPound= (float)(A+(B*V)+(C*Math.pow(V,2))+(K*D*Math.pow(V,3)));
walkingBurnedCalories= ratePerPound*weightInPounds;
WalkingBurnedCalories.setText ("Walking burned calories : "+ walkingBurnedCalories + "
calories \n");
}
```

Get weather temperature: The following steps show, how weather information is collected.

- To acquire data from temperature and humidity sensor, an instance of the SensorManager class is created. This instance is used to get the physical sensor.
- Register a sensor listener in the onResume() method, and start handling incoming sensor data in the onSensorChanged() callback method.
- Implement onAccuracyChanged() and onSensorChanged() callback methods. The sensor is unregistered when an activity pauses to prevent the sensor form continually sensing data and draining the battery.

The code shown in Appendix 1(b) uses the temperature and humidity sensor to get the weather, room temperature and then display it for the user.

Share pictures: In order to share pictures, the applications issues a HTTP POST request to share a photo with friends or healthcare professionals.

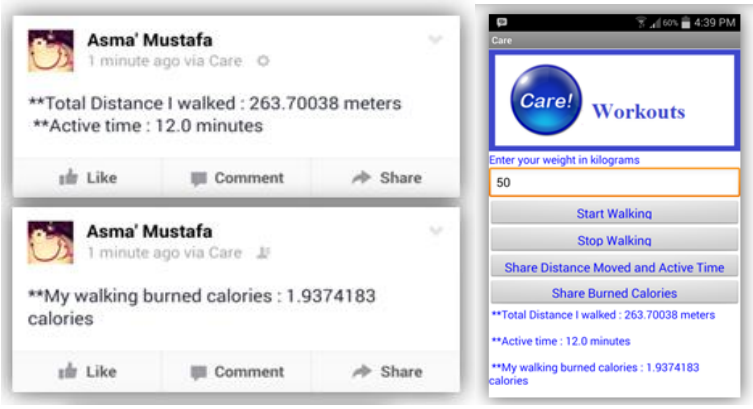


Figure 6: The application workouts

Hear rate monitor: Heart rate monitor uses the camera and its flash to find the user heart rate in beats per minute. The user has to hold the tip of his/her index finger over the camera lens of his/her phone. The application takes between ten to thirty seconds to get accurate heart rate. Heart rate monitoring is based on using the camera with as little focus as possible. When the user puts his/her finger on the camera lens, it won't be focused. The resulted image will be only shades of light and dark RGB. The code looks at single channel (red) and tries to find out when the channel goes from light to dark red.

The application uses the 'PreviewCallback' mechanism to capture the last image from the preview frame. Then the YUV420SP data will be processed to get all the red pixel values. Data smoothing in an integer array is used to figure out the red pixel average value in the image. The heart beat is detected when the average red pixel value in the latest image is greater than the smoothed average.

The application collects data during ten seconds, and then adds the beats per minute to an integer array which will be used to smooth the beats per minute data. As an illustration, the Figure 7 is the resulting display of user heart rate using this application and another medical device at the same time. It turned out that the results are similar.

Implementing Privacy Model: In order to maintain privacy, the application provides the user with three roles: trusted, semi-trusted, and un-trusted as shown in Figure 8. Predefined permissions are assigned to each one of the roles. The members of each role will be assigned by the user from his contact list.

When the user presses Pick Trusted Friends, the following steps will occur:

Issue a HTTP GET request, shown below, to /fql?q=query to get user friends list as shown in Figure 9:

```
String query = "select name, current_location, uid, pic_square from user where uid
in (select uid2 from friend where uid1=me()) order by name";
Bundle params = new Bundle();
params.putString("method", "fql.query");
params.putString("query", query);Utility.mAsyncRunner.request(null, params, new
TrustedFriendsRequestListener());
```

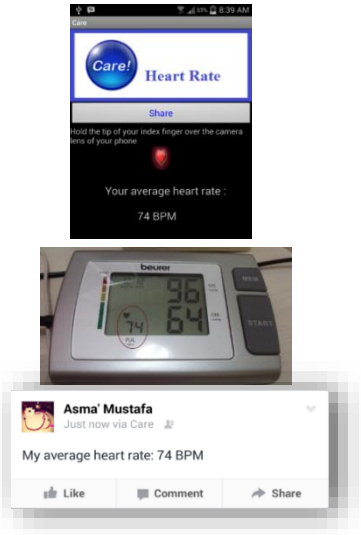


Figure 7: User shared heart rate

Figure 8: Privacy in the application

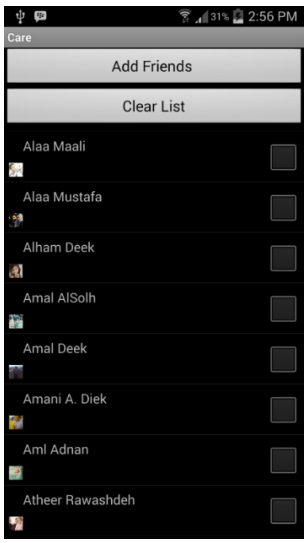


Figure 9: User friend list

When the user press Add Friends, after choosing the ones he wants to select as trusted, the following steps will occur:

- Issue a HTTP POST request to /fql?q=query to create Trusted Friend list.
- Issue a HTTP GET request, shown below, to /fql?q=query to get Trusted Friend list id, then, a POST request to add the selected friends to the list

The same process will be executed when the user presses Pick Semi-Trusted Friends or Pick Un-Trusted Friends.

When the user presses Assign Trused Relationships, the user will get the layout as shown in Figure 10. If he/she pressed 'Close Friends' or 'Un Close Friends', he/she will get a list of the trusted friends only to choose the close or not close friends from them.

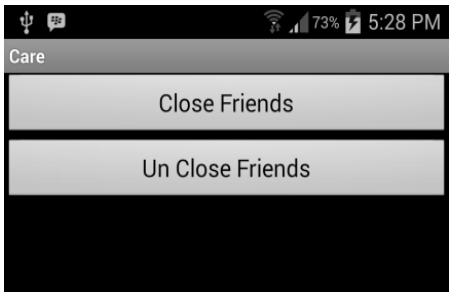


Figure 10: Assigning trusted relationship

When the user presses 'Close Friends', the following steps will occur:

1. Issue a HTTP GET request to /fql?q=query to get user trusted friends only as shown in Figure 11:

```
graph_or_fql = "fql";
String query = "select name, current_location, uid, pic_square from user where uid
in (SELECT uid FROM friendlist_member WHERE flid ="+ trustedFriendListId +") order
by name";
Bundle params = new Bundle();
params.putString("method", "fql.query");
params.putString("query", query);
Utility.mAsyncRunner.request(null, params,
new assignCloseTrustedRequestListener());
```

Adding Friends to close or un-close will be performed the same way as mentioned in the trusted friend list.

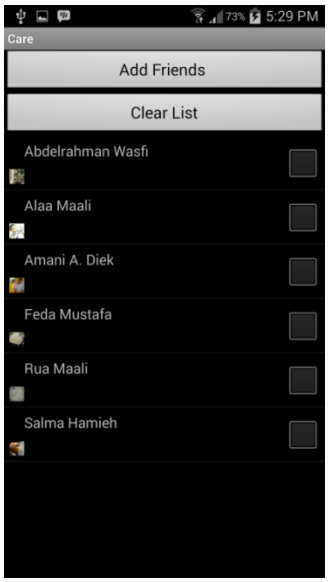


Figure 11: Trusted friend list

Each time a post will be shared such as 'distance moved', 'calories burned', 'active time', 'weather temperature', 'heart rate or pictures', the following steps will be performed to specify who can see and comment on the post and who can't. Below is an example for sharing weather temperature. As we can see trusted and semi trusted friends are allowed to see and comment on the weather post while untrusted friends are denied.

```
private void share() {
String s="";
Bundle params = new Bundle();
params.putString("message", temperature);
JSONObject privacy = new JSONObject();
try {
privacy.put("value", "CUSTOM");
privacy.put("friends", "SOME_FRIENDS");
privacy.put("allow", trustedListId+","+semiTrustedListId);
privacy.put("deny", unTrustedListId);
} catch (JSONException e) {
 // TODO Auto-generated catch block
e.printStackTrace();
}
params.putString("privacy", privacy.toString());
try {
Utility.mFacebook.request("me/feed", params, "POST");
Toast toast;
toast = Toast.makeText(Temperature.this,"Your status has been updated",
Toast.LENGTH_LONG);
toast.show();
} catch (FileNotFoundException e) {
// TODO Auto-generated catch block
e.printStackTrace();
} catch (MalformedURLException e) {
// TODO Auto-generated catch block
e.printStackTrace();
} catch (IOException e) {
// TODO Auto-generated catch block
e.printStackTrace();}
```

5 Comparative Analysis

A number of recent applications designed in context of integrating wireless sensor networks with online social networks can be examined for the purpose of comparison. The existing platform applications such as Google Latitude shares the collected mobile position data of the user among different users, and then it generate proximity alerts when two linked users are within geographical proximity of one another. These applications are limited and target specific service only. In this work, the application not only uses the built-in sensors of the user mobile device (i.e., get current location coordinates, view user current location, share location, show nearby places, and show nearby friends), but also provides more services such as clustering nearby friends, tracking distance moved, calories burned and active time, hear rate monitoring, getting weather temperature and sharing pictures. This may help to improve healthcare awareness and prompt quicker and safe advice from healthcare professional in social network. In this work, the data privacy is enforced by both options: the first option is available on all available social network platforms – as ON/OFF, while the second option is using user assigned roles versus data levels.

6 Conclusions and Future Work

In this paper, the framework and implementation of an application was presented. A number of sensors were used to build a sensing application that enables members of a social network to share their information with their contacts in a private manner. The user can see if any of his/her contacts are checked-in nearby. It was shown that contacts can be clustered based on colleagues, family or any other criterion. The clustering process takes place in the user device. The application is expected to be a great tool for fitness, weight loss, calorie counting, etc., and can facilitate quicker monitoring of, for example, heartbeat of the user through social network by concerned healthcare units. Social constraints such as privacy were addressed in two ways: simple feature like turn application ON or OFF; and the other by privacy aware data connectivity based on user roles.

Some future work can be performed in order to extend and improve the built application. Some of its aspects can be improved and more functionality added. Below are some suggestions for further improvement:

- Health support for elders by using sensor information to send alerts if there is abnormal activity. Request for attention can be sent to doctors and nearby friends based on the collected information from sensors such as body position and health measurements.
- Suggesting nearby friends based on common interests.
- Adding more features such as monitoring and tracking user blood pressure. This also includes storing, analyzing and sharing the user blood pressure measurements.

REFERENCES

[1]. M. Blackstock, R. Lea, and A. Friday, “Uniting online social networks with places and things,” in Proc. of the Second International Workshop on Web of Things, New York, NY, USA, 2011, pp. 5:1–5:6.

[2]. C. Aggarwal, T.F. Abdelzaher, “Integrating Sensors and Social Networks”, Social Network Data Analytics, chapter 14, pp. 379-412, Springer, 2011.

[3]. E. Miluzzo, N. D. Lane, S. B. Eisenman, A. T. Campbell, “CenceMe: Injecting Sensing Presence into Social Network Applications using Mobile Phones”, in Proc. of the 2nd European Conference on Smart Sensing and Context, Springer, October, 2007, pp. 1-28.

[4]. R. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. Abdelzaher, “GreenGPS: A Participatory Sensing Fuel-Efficient Maps Application”, in Proc. of Mobisys, San Francisco, CA, June 2010, pp. 151-164.

[5]. Q. Memon, S. Khoja, “Academic Program Administration via Semantic Web – A Case Study”, Proceedings of International Conference on Electrical, Computer, and Systems Science and Engineering, Dubai, Volume 37, pp. 695-698, 2009.

[6]. Q. Memon, S. A. Khoja, “Semantic Web Approach for Program Assessment”, International Journal of Engineering Education, Vol. 25, No. 5, pp. 1020-1028, 2009

[7]. A. Moravejosharieh, H. Modares, R. Salleh, "Overview of Mobile IPv6 Security," in Proc. of 3rd International Conference on Intelligent Systems, Modelling and Simulation, 2012, pp. 584-587, DDI: 10.1109/ISMS.2012.9.

[8]. Q. Memon, "A New Approach to Video Security over Networks", International Journal of Computer Applications in Technology, Vol. 25, 2006, pp. 72-83.

[9]. Y. Altshuler, Y. Elovici, N. Aharony, and A. Pentland, Security and Privacy in Social Networks, Springer, 2012.

[10]. M. Huber, M. Mulazzani, E. Weippl, G. Kitzler, and S. Goluch, "Exploiting social networking sites for spam," in Proc. of 17th ACM Conference on Computer and Communications Security, NY, USA, 2010, pp. 693–695.

[11]. B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, (CSUR), Vol. 42, No. 4, 2010, doi: 10.1145/1749603.1749605

[12]. Q. Memon, S. A. Khoja, "RFID based Patient Tracking in Regional Collaborative Healthcare", International Journal of Computer Applications in Technology, Vol. 45, No. 4, 2012, pp. 231–244, doi: 10.1504/IJCAT.2012.051123

[13]. Q. Memon, S Akhtar, AA Aly, "Role management in adhoc networks," Proceedings of the Spring Simulaiton Multi conference-Vol.1, 2007, pp. 131-137.

[14]. K. M. Karkanen, "Walking/running Heart Rate Monitoring System," U.S. Patent 6013009, Jan 11, 2000.

TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

A Comparative Analysis of Privacy Preserving Techniques in Online Social Networks

[1]Firdous Kausar and [2]Shoroq Odah Al Beladi
Department of Computer Science, College of Computer and Information Sciences,
Al Imam Mohammad ibn Islamic Saud University, Saudi Arabia
[1]firdous.kausar@ccis.imamu.edu.sa, [2]shorogodah@gmail.com

ABSTRACT

The world became a global village due to the great revolution in communication and network fields. The internet eases the communication process all over the world. The social network services, for example; Facebook, that occurred recently are considered one of the most essential and common outcomes due to this revolution. People from different ages, views, cultures, languages, religions, education levels, etc. from all over the world can easily communicate through the social networks. Furthermore; people can share their news and personal information with each other. It can be also noticed that the social network are extensively used, so the privacy issues within it is essentially to be considered. This privacy will prevent any illegal accessing for the user personal information which in turns will increase the users' conviction to use these networks during their daily life and encourage the other people who have never use the services that are available for using by social networks. This paper analyses several techniques which provides the privacy for Online Social Networks including Reclaim, Safebook, K-Automorphism, Vis-à-vis and SPKA. These techniques are investigated and clarified in terms of their methods in addition to the advantages and drawbacks.

Keywords: Online Social Networks (OSNs), Privacy, Communication and Transport (CT), Social Networking Services (SNSs), Distributed Hash Table (DHT), Virtual Individual Server (VIS), Trusted Identification Service (TIS).

1 Introduction

The great revolution that emerged recently because of lead to a clear development in all life filed. This development is required to enable the users all over the world performing their tasks and works in addition to enhance the communication tools and skills among the users all around the world. Due to this revolution; "Online Social Networks (OSNs)", including Facebook, Twitter and Google +, emerged as a common and deployed communication tools that can be used for sharing the information between these users. These networks have attracted a very large number of users. Furthermore; the adoption is still increasing day by day [1].

Several services are provided to the users due to using OSNs, such as; instant messages, internet phones and blogging without giving considerations for the physical location of the users. Furthermore; the friends and families could easily maintain their relations in more reliable and convenient manner in comparison with old style contact techniques, such as; phone conversation and emails. These networks in turns also increase the popularity for the users within social groups; all these are considered attractive advantages for OSNs. On the other hand; there is some drawbacks for OSNs that limits the deployment of it and always increase the motivation for new techniques

DOI: 10.14738/tnc.32.854
Publication Date: 20th April, 2015
URL: http://dx.doi.org/10.14738/tnc.32.854

that will shrink and eliminate the effect of these drawbacks, the privacy concern is considered one of the most critical and significant issue within OSNs. The privacy risks increases in case that the personal information for the users, such as names; are included within OSNs applications. As a result; there is an increasing demand for novel privacy schemes that will in turns preserve the users' personal information and ensure their privacy [2].

The privacy is considered one of the OSNs security objectives in addition to the availability and integrity [12]. Maintain the privacy of the users is the most essential and critical objective for "Social Networking Services (SNSs)". The privacy issue is concerned in personal information protection that is shared and published on the profiles of users in addition to maintaining the privacy during communication [11]. So; only trusted parties are able to trace the communicating parties. In addition to that; the privacy is also concerned in hiding the message details in such way that only the receiver and sender can recognize it. To conclude all; the privacy is concerned in personal information hiding for the users. The privacy should be achieved by default [3].

The breaches of privacy within social network can be classified into three main groups, which are [4];

- Identity detection; this type of breach occurred in case that individuals that the record belong to is disclosed. This will in turns result in information revelation of users and the shared relationship from them with other network's individuals.
- Sensitive link detection; this type of breach occurred in case of the association revelation among two users. This information is generated by social activities in case of utilizing the services of social media by the users.
- Sensitive attribute detection; this type occurred in case that the confidential and sensitive user information is being attained by the attacker. Sensitive attributes are related to the link and entity relationship.

Three levels are included within SNSs as illustrated below in Figure 1 below [3].

Three levels are included within SNSs as illustrated below in Figure 1 below [3].

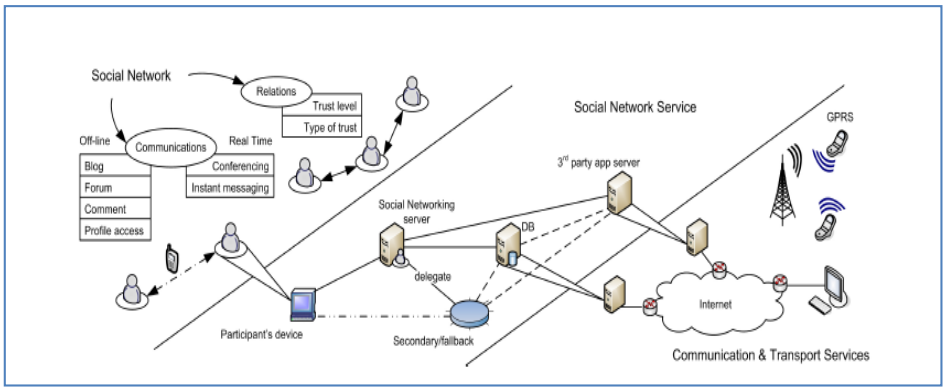


Figure 1: levels of SNSs [3]

As illustrated in figure 1; the SNSs levels can be summarized as listed below;

- The level of "Social Network (SN); this level includes the members relationship digital representation.
- The level of SNS; in this level; the infrastructure of the application that is controlled through SNS provider.
- The level of "Communication and Transport (CT)", this level includes the transport and communication services that are provided through the network.

During SN level; the members are provided with several functions regarding to the real life interaction, such as; profile accessing, like, finding friends and commenting. All these functions are implemented relying on the level of SNS. SNS level in turns includes the main services that are required or needed for generating SN services, including, storage, web services and communication. The delegation and redundancy are considered two of the most commonly deployed techniques for availability improvement [3].

Data retrieval and storage, content indexing, data access permission management in addition to the leave or join of node, are all implemented using either decentralized or centralized fashion of distribution within SNS level. Furthermore; the internetworking infrastructures and protocols that were previously implemented within CT level are used during SNS level. Depending on OSN architecture; the attacker can be defined as of the following type [3];

- SN level malicious member
- SNS level service provider.
- CT level malicious that have illegal access for the infrastructure.

One main aspect that must be achieved within OSNs is the privacy issue; it is mainly concerned in protecting the identity of the members. For example, the identity theft is defined as malicious service provider or member obtains the authorized users credentials and then take action rather than them by getting access for the profiles of these members. Furthermore; the intrinsic trust among people also plays a great role in privacy violation for the users via making another copy of the target personal profile using the personal information and then start communicate with others via sending them friend request for example. Another type of probable attacks is the profile porting, in this type; the target person profile is created in OSN by the attacker without present of the victim. The detection for this type of attack is usually difficult mission [3].

Several research papers and studies investigated and evaluated the privacy issue within OSNs. Generally; the privacy can be achieved through network decentralization or encrypting the data before storing it. The stimulants of reading the private data is the condition for sustain the centralized network. In decentralized type of network, there is no incentive that the network relies on, so; the operation of the network can be continued depending on the contributed resources of the user [5].

2 Privacy Preserving Techniques

In this section we provide the description of different privacy preserving techniques.

2.1 Cachet

The cachet [1] is a structural design by which a strong privacy and security can be achieved for the users maintaining the OSNs main functionality. The availability, confidentiality and integrity in addition to the users' relationship privacy can be protected using cachet. The user data is stored using distributed nodes pool; this pool is also used for availability ensuring [14]. Since the cachet storage nodes are not trusted; then a technique of leverage cryptographic, specifically; "Attribute-Based Encryption (ABE)", is employed in order to achieve the trust for the storage nodes within cachet, which in turns will protect the data confidentiality [13]. Furthermore; "Hybrid Structured-Unstructured Overlay Paradigm (HSUP)" is also employed to achieve efficient retrieval and dissemination for the data. So; a "Distributed Hash Table (DHT)" is augmented regarding to the users social links. In order to minimize the overhead on the network, for example; reducing cryptographic, then the recent updates within SN is stored by the social contacts that operate as caches. FreePastry

Simulator was employed in implementing the cachet prototype. Furthermore; newsfeed application was also implemented to illustrate the existing OSNs functionality. An example for this technique is illustrated below in the following figure 2

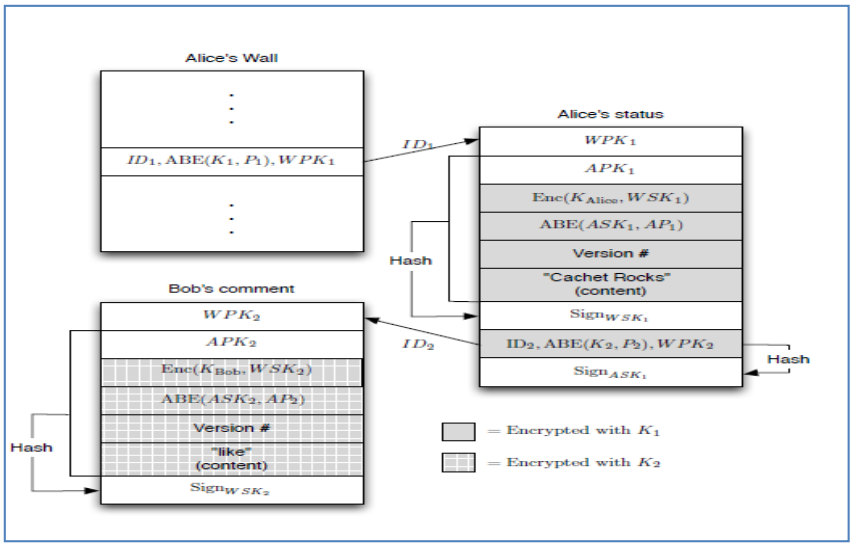


Figure 2: Cachet Example [1]

For Alice to be joined to the cachet; the varouis keys should be firstly generated, a wall and profile inforamation is created and finally these inforamation are saved as wall and root information within DHT respectively. In order to set up a friend or co-worker relationship with Bob, then an a secret key is generated for Bob using ABE among the co-worker and friend attributes. A different realtionship may be established by Bob among Alice. If Alice wants to post a status update, then a status object is created by alice, in addittion to the number of version, content and secret and public keys for appended polices and write (WPK1,WSK1,APK1, ASK1). “Write Ploicy Sgnature Key (WSK1)” is used in signitaure genreation. A symettric key for encryption say K1 is randomly picked by Alice and use it in encrypting the objects (excluding APK1, WPK1 and signature) [1].

Furthermore; an ID say ID1 is also selected randomly and then used in inserting the objects within DHT. A refreence is finally created for the update of the status including K1, ID1, WPK1 and then added to the wall. This update can be read by Bob through finding referrence on the wall of Alice, after that the attribute-based key that was prevouisly transmitted by Alice for alice will be used in decrypting K1. The object is then retrieved from DHT in additiion to ID1, the encrypted field is then decrypted employing K1. The object integrity is then being ensured through signature verification. If Bob wants to comment on the status update of Alice; then the same steps that were followed by Alice during creating the update will be followed in creating the comment object. Append operation is then used in referrence inserting realted to the novel object into the update of Alice. If AP1 is staisfied, then ASK1 is decrypted to be latterly employed in signature geneartaion.

2.2 ReClaim

ReClaim[5] is a solution for decntralized OSNs, each one of the peers is named as Reclaim. friends of friend are emplyed in replicating the data, so; no addittional server is required for storing the data. The public keys are exchanged in order to set up the frindship, these keys are in tuns employed in encrypting all messages and make them redable for only the target user. “Private-Set-Intersection (PSI)” approch [10] is employed in the current online friends. So; two peers are allowed to discover the common friends without disclosing the unmutual ones. Missed or older messages are then

syncronized to common friends by the two peers using Bloom filters [9]. the encryption for all messages that were prepared for the peers and their friends can be performed effciently, furthermore; the duplicate transfers of messages is prevented. By this way; the operation of ReClaim can be reliably continued despite of intermittent connection of the network, the firewall of "Network Address Translation (NAT)" and network dealy. A method that is recognized as $FSF_{A,B}$ is employed in detrmining whether that the just connected peers are either friend or hanve common friends. The SFS inputs is denoted by $\mathbb{F}$ and it is recognized as friendset that includes the user and the friends identifiers.

2.3 Vis-à-vis

The concept for this decentralized framework is summarized in maintaining the privacy for the "Virtual Individual Server (VIS)". In this technique, each person data is stored within him/her VIS. Vis-à-vis is a privacy technique that is concerned in location information privacy. The information location can be efficiently shared within groups through employing the trees of distributed location. The architecture for Vis-à-vis is illustrated below in the following figure 3.

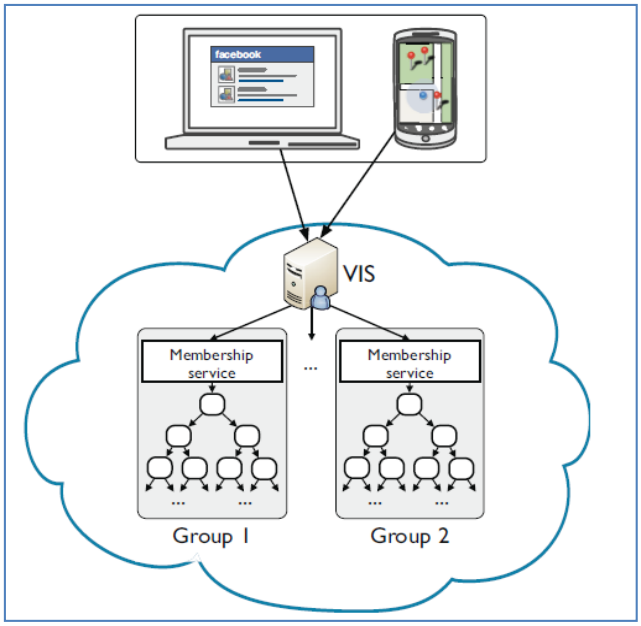


Figure 3: architecture of Vis-à-Vis [6]

The process of naming the groups is performed through employing a descriptor which in turns includes the public key related to the group in addition to specific string used to transfer the shared attribute between the members of the group. The expression for the descriptor is given by$< K+_{owner}, string >$; $K+_{owner}$ represents public key. No sensitive information in terms of the privacy is included within descriptors, for instance; the public key for the user is shared among the group. As shown in Figure 3; the architecture of Vis-à-Vis looks like the structure of distributed tree. The hierarchical structure is adopted to be used over DHT; since the required range quires for search operation is not easily provided when applying DHT. The groups that based on the location are accessed by the users throughout clients, for examples; web browsers and mobile applications. Vis-à-vis was designed to deal with established OSNs, for example; Facebook. A stronger privacy is achieved for the users when employing Vis-à-Vis in comparison with centralized services, for examples; MySpace and Facebook. This happens because the users are provided with more control for their personal data and who can access it. The trust model is designed based on the compute utilities business interest in addition to the user's social relationships [6].

2.4 Safebook

Safebook is an architecture in which three tiers are included in addition to the layers direct mapping with OSNs level that were previously introduced, this architecture is illustrated below in the following figure 4.

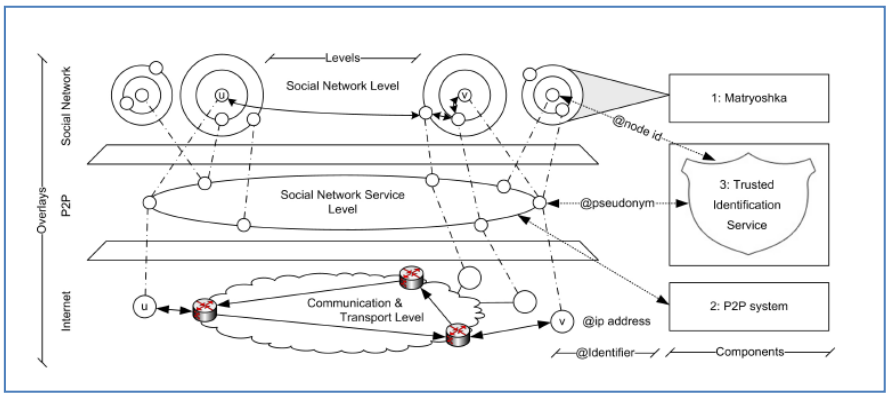


Figure 4: Safebook architecture [3]

From Figure 4; it can be concluded that;

- CT level is represented by Internet.
- SNS service is implemented using "Peer to Peer (P2P)" base.
- SN level is implemented using user-centered layer of social network.

A node is used to present Each Safebook party. In internet, this node is host node, in case of P2P overlay then it is peer node and member within SN layer. Two types of overlay are formed by the Safebook nodes, which are [3];

- P2P base or substrate; this type is concerned in lookup service providing.
- Matryoshkas set; this type is considered SN concentric structure and it concerned in data storage providing in addition to the creation of communication privacy around all nodes. The structure of Matryoshkas is illustrated below in the figure 5.

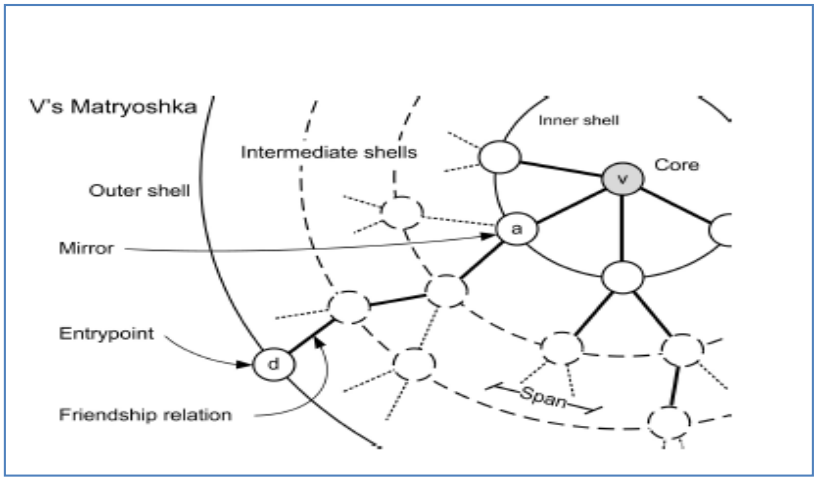


Figure 5: Structure of Matryoshkas [3]

Furthermore; a "Trusted Identification Service (TIS)" is alfo featured by safebook in order to give an unambigous identified for each node; the Pseudonym in addittion to the "Node Identifier (NI)". Particular countermeasures set is implemented by each component in Safebook in order to

overcome the probable threats in OSNs. As shown in Figure 5; each single Matryoshkas is concerned in providing the protction for the core node that is addressed using NI within SN layer. Radial paths are used to connect the nodes; so the message can be recursively transmitted among shells. The trust relationship that is the most similar to social network is considered by the paths; so a hop is used to connect two nodes related to users who have real-life trust relationship. The mirrors are the nodes that have direct contact with core; the data are stored in these nodes in encrypted form. The remaining nodes excluding the core and mirrors are recognized as entrypoints which in turns act as gateway for passing all request of data to the core node [3].

2.5 K-Automorphism

K-Automorphism is also a privacy technique for OSNs; it is concerned in the problem of identity disclosure through employing "K-Match (KM)" technique. Table 1 summarizes illustration for the main symbols that are employed within this algorithm [7].

Table 1: K-Automorphism symbols [7]

Symbol	Description
G	Main network
$G^{'}$	Gullible anonymized network
$G^{''}$	Anonymized graph with alignment and portion
G^{*}	Anonymized graph with KM technique
$\overline{G^{*}}_{t}$	Anonymized graph with GenID technique at T_t time.
U_i	Blocks group.
P_{ij}	A single block within U_i group

The privacy within K-Automorphism is provided against structural attack as summarized as follow; for network G; if a $k-1$ numbers of automorphic functions F_a $(a = 1, 2, \ldots\ldots, k-1)$ are defined within G, and the relation $F_{a2}(v) \neq F_{a2}(v)$ for each v where v is a vertex in G, then the K-automorphic network take place. The vertex v cannot be differentiated from other similar vertex depending on the structural information [7]. The privacy in this technique is achieved through using KM technique that is summarized below in the figure 6.

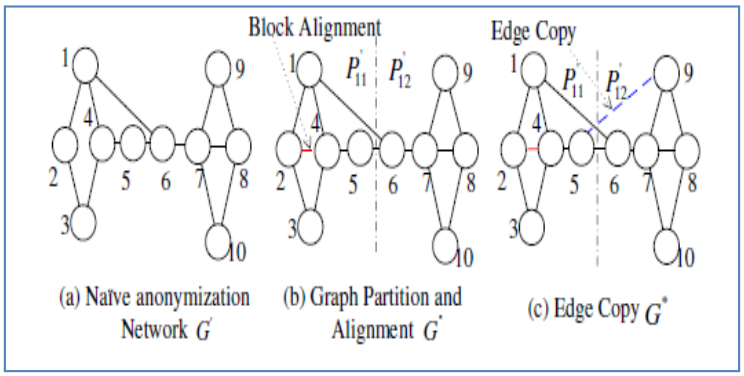


Figure 6: KM technique [7]

As shown in figure 6; if 2-different matches are required to be guaranteed within network G, then g is firstly divided into two blocks (Figure 6b). The graph alignment is then performed to attain two aligned blocks. In case that query Q match is included within G^{*}, then 2 matches are least exist of Q within G^{*}. The principle of k-different match is still able to be applied after performing the alignment and partition for the network [7].

2.6 Single Pass K-means Anonymization (SPKA)

This technique mainly based on the method of K-mean clustering by including only one iteration rather than multiple ones. If D represents the records set that will be anonymized, K represents the clusters' number and n represent the records' number. The SPKA technique can be summarized as follow. The quasi-identifiers are used to classify the records that are within D. a K records then randomly selected to represent the early cluster that will used in generating all later K clusters. For $r \in D$; r is assigned to the center of nearest cluster. The centers of the clusters are updated in case of adding novel records. In order to decrease the losses of information, some adjustments are then performed. Some records are taken away from the clusters that consist of number of records that is larger than k and then added to these clusters that consist of number of records that is less than k. in case of not having clusters with number of records less than k; the removed records are then allocated within particular closest clusters. As a result; a complexity of $O(n^2/k)$ is achieved. The t-closeness privacy measure was employed during their investigation. In this technique; the distance among the sensitive attribute distributions and the total table attribute distribution is required to be less than predefined threshold value. "Earth Mover's Distance (EMD)" is employed in computing the distance among two distributions.

3 Comparative Analysis

The privacy and security are considered two of the most essential and critical concerns in social network. Several techniques can be employed to achieved these two concerns, Six techniques were introduced and investigated during this research, which are; Reclaim, Cachet, Vis-à-vis, Safebook, K-Automorphism and SPKA. Each one of the considered technique has it is own concepts and methodology in order to provide the social network users with the required level of privacy, which in turns result in protect their personal information from being accessed by unauthorized persons.

The first considered technique is Reclaim technique; this technique is mainly based on employing PSI, $\mathrm{FSF}_{\mathrm{A,B}}$ method and Bloom Filters in order to provide achieve users' privacy. This technique has several advantages that make preferable to be used and employed in social network. The first point is the decentralized architecture; which in turns means that the communication can be performed between peers and users without need for any central equipment. Furthermore; this techniques has the advantages of low running costs, able to deal with extreme churn, overcoming NAT firewall, enabling the communication among the friends despite of being offline and accessing the replicas easily from the users. On the other hand; the complexity within Reclaim structure may limit the deployment for this technique.

The second considered technique is Cachet technique; which mainly implemented based on Nodes distributed pool, ABE, HSUP concepts in addition to Social contact employment. High level of security and privacy is provided to OSNs users. Furthermore, this technique provides the users with availability, confidentiality and integrity protection. On the contrary; this technique essentially still needs Pliability against the churn of node. Furthermore; there is difficulty in recognizing the peer to peer connection for the users who are located at the NAT back. This technique also required additional bandwidth, computational resources and data storage.

The third technique is Vis-à-vis technique. This technique based on VIS and Distributed trees concepts. Decentralized framework is also considered in this technique in addition to Hierarchical structure. This technique was designed to achieve privacy on location information only and some improvement and development is still required to be added to the design in order to include other information privacy. Furthermore; the breaches cannot be totally eliminated.

The fourth technique is the Safebook; which mainly based on Matryoshkas, DHT and Real-life trust concepts in order to provide the OSNs users with the required level of privacy and security. The decentralized structure is also considered here. Furthermore; Feasible Realistic compromise among performance and privacy is also provided. Extra costs are required in case of using additional hobs to enhance the privacy.

The fifth considered technique is K-Automorphism, which based on KM algorithm, Edge copy, Graph alignment and Graph Partitioning concepts. The structural attack can be got over using this technique. Another advantage for K-Automorphism is that No uncertainty for the released network. The privacy can be also provided in dynamic release case. The complexity in finding automorphic functions may limits the deployment for this technique.

The last technique is SPKA that is mainly based on Clustering and T-closeness approaches. Similarity attack can be avoided using this technique. It also achieves better privacy level in comparison with l-diversity technique. An additional benefit is the Quasi-identifier data prevention. The complex process and the added overhead for anonymization process are two limitations for SPKA technique.

The table2 introduces a comparison between the considered privacy techniques for OSNs. The comparison is held based on the main concepts that were employed during the implementation and investigation of the techniques in addition to the benefits that encourage the use of these techniques in OSNs services and applications. The drawbacks that limit the deployment of them are also concluded.

Table 2: Comparison between different techniques

Technique	Main concepts	Strength/Benefits	Weakness/Challenges
Reclaim	-PSI -$FSF_{A,B}$ method -Bloom Filters	-Decentralized architecture. -Very low running costs. -Ability to deal with extreme churn. -Get over NAT firewalls. -Enable the friend communicating despite of being offline. - Easy replicas accessing from the users.	-Complex Structure.
Cachet	-Nodes distributed pool. -ABE -HSUP Social contact employment.	-Strong privacy and security in addition to maintaining OSNs main functions. -provide availability, confidentiality and integrity protection in addition to preserving privacy. -Practical ABE_decryption computational overhead.	- Still essentially requires pliability against the churn of node. -there is a difficulty in recognizing the P2P connection for the users who are at the back of NAT. -Extra computational resources. - Volunteering bandwidth and data storage.

Vis-a-Vis	-VIS. -Distributed trees.	-Decentralized framework. -Hierarchical structure.	-supply the users with only location information Privacy and some improvement is still achieve the privacy for other data types. - The breaches cannot be totally eliminated
Safebook	-Matryoshkas. -DHT -Real-life trust	-Decentralized structure. - Feasible Realistic compromise among performance and privacy.	-Additional costs is needed for increasing privacy be adding additional hops.
K-Automorphism	-KM algorithm. -Edge copy -Graph alignment -Graph Partitioning	-Can overcome structural attack. -No uncertainty for the released network. -Provide Privacy in case of dynamic releases.	-Complex process for finding automorphic functions.
SPKA	-Clustering. -T-closeness	- Quasi-identifier data prevention. -Outperforms the privacy of l-diversity technique. - provide privacy against similarity attack.	-Overhead due to applying anonymization process. -Complex technique.

4 Conclusion

The OSNs are now widely deployed and used by large number of users from all ages and different views all over the world. The security within these networks is considered essential concern that must be achieved in order to maintain the personal information of users from being accessed or recognized by un-authenticated users or attackers. The privacy is one of the security objectives within OSNs in a line with the integrity and the availability. This paper introduced six common techniques used in preserving the privacy within OSNs. A brief explanation for the techniques was introduced in addition to investigation of the main concepts that were employed to implement the technique. As a result of comparative analysis between the different privacy preserving techniques it has been found that each technique has its own potential benefits and drawback. Some of the drawback of these techniques include 1) have complex architecture 2) require excessive computational recourse 3) need special hardware and 4) not simple to implement. On the other hand these also provide a strong benchmark for providing anonymity and privacy in online social networks.

REFERENCES

[1] Nilizadeh, S, Jahid, S and Mittal, P, “Cachet: A Decentralized Architecture for Privacy”, the 8th ACM international Conference on emerging networking experiments and technology, 2012.

[2] Sun, J, Zhu¢Ó, X and Yuguang Fang, Y, “A Privacy-Preserving Scheme for Online Social Networks with Efficient Revocation”, IEEE INFOCOM, 2010.

[3] Cutillo, L, Molva, R, Strufey, T and Eurécom, I, “Safebook: a Privacy Preserving Online Social Network Leveraging on Real-Life Trust”, 2007.

[4] Vijayalakshmi, V, Arunachalam, A and Nandhakumar, R. “Mining Social Media-Utility Based Privacy”, (IJCSIT) International Journal of Computer Science and Information Technologies, 5 (4), 5480-5485, 2014.

[5] Zeilemaker, N and Pouwelse, J. “ReClaim: a Privacy-Preserving Decentralized Social Network”, 4th USENIX Workshop on Free and Open Communications on the Internet, August 2014.

[6] Shakimov, A, Lim, H, C´aceres, R, Cox, Li, P, Liu, D and Varshavsky, A. Vis-`a-Vis: Privacy-Preserving Online Social Networking via Virtual Individual Servers. IEEE, 2011.

[7] Zou, L, Chen, L and Ozsu, M, “KAutomorphism: A General Framework for Privacy Preserving Network Publication”, 2009.

[8] Poulin,I and Kani, M, “Preserving the Privacy on Social Networks by Clustering Based Anonymization”, International Journal of Advanced Research in Computer Science & Technology (IJARCST), 2 (1), pp.11-14, Jan-March 2014.

[9] Bloom B. H. Space/Time Trade-Offs nn Hash Coding with Allowable Errors, Communications of the ACM 13, pp. 422–426, July 1970.

[10] Freedman, M. J., Nissim, K., And Pinkas, B. Efficient Private Matching and Set Intersection, in EUROCRYPT ’04, pp. 1–19, 2014.

[11] Beato, F., Kohlweiss, M., and Wouters, K, Scramble! Your Social Network Data, in PETS ’11, vol. 6794 of Lecture Notes in Computer Science, pp. 211–225, 2011.

[12] Elena Z, Lise G, Privacy in Social Networks: A Survey, in Social Network Data Analytics, pp 277-306, 2011.

[13] J. Bethencourt, A. Sahai, and B. Waters. Ciphertext-policy attribute-based encryption. In IEEE Security & Privacy, 2007.

[14] S. Jahid, S. Nilizadeh, P. Mittal, N. Borisov, and A. Kapadia. DECENT: A decentralized architecture for enforcing privacy in online social networks. In SESOC, 2012.

TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

Reversible Nanoporous Sensors of Carbon Monoxide in Atmosphere

Alexander Novikov
Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Physical Engineering, Kronwerksky, Saint Petersburg, Russia.
afnovikov@mail.ru

ABSTRACT

Optical chemical compositions being sensitive to carbon monoxide in atmosphere were investigated and tested. The gas sensitive Palladium (II) complexes were immobilized within nano-scale through-out porous glass substrates. Fabricated specimens have demonstrated reversible color changes while interaction with gaseous carbon monoxide in atmosphere.

Keywords: Optical chemical sensors, Carbon monoxide detection, Nanoporous glass composition materials, Palladium complexes.

1 Introduction

Atmosphere content analysis often requires the portable sensors of carbon monoxide (CO). The most popular CO detectors are the single-shot Dräger indicator tubes [1] using particularly the iodine chemical reduction by CO. Also a row of various sensor devices intended for CO detection are now at disposal. They are based on different principles such as electrochemical, catalytic, optical and infrared absorption, etc. sensors [2-5]. In general, those sensors demonstrate enough good performances, however, accompanied with certain disadvantages mainly connected with so-called cross-sensitivity. As a rule, the designers have to solve a problem of distinguishing namely CO component in presence of other reducing atmosphere components. So an attempt to find out new solution of the said problem seems to be just relevant.

Specific chemical reaction of the CO gas with Palladium (II) salt is well-known [6]. During a solution barbotage with gas, the metallic Palladium rapidly precipitates, so that originally brown solution instantly and irreversibly gets colorless:

$$PdCl_2 + H_2O + CO \rightarrow Pd\downarrow + CO_2 + H_2,$$

or

$$[PdCl_4]^{2-} + H_2O + CO \rightarrow Pd\downarrow + CO_2 + 2H^+ + 4Cl^-.$$

The latter reaction supposedly proceeds in two steps as follows [7]:

$$[PdCl_4]^{2-} \xrightarrow{+CO} [PdCOCl_3]^- \xrightarrow{+H_2O} Pd\downarrow + CO_2 + 2H^+ + 4Cl^-$$

Principally, the process can't be run gradually, and a required sensing instrument calibration is impossible. However, the time spreading the process may be provided with use of the more complicated solution composition and procedure as shown in [8].

DOI: 10.14738/tnc.32.899
Publication Date: 20th April, 2015
URL: http://dx.doi.org/10.14738/tnc.32.899

But probably the most prospective way in this sense were to abandon the liquid medium at all and so to immobilize certain palladium complexes on a dry dielectric surface.

This seems to be available with use of the certain palladium complexes being adsorbed on the properly pretreated dielectric surface. As an appropriate substrate in similar applications proved to be a matrix made of the nano-scale porous silicate glass. Presently, the structure and properties of nanoporous glasses as they are have been thoroughly studied by many researchers [9-11].

Methods for precipitation of various dopants within nano-scale porous silicate glass have been invented enough long time ago [12]. However, only during recent decades the intensive studies have been carried out having as an object the organic and inorganic molecular clusters down to single molecules being immobilized on the highly developed surface of the nanoporous glass [9,10,13].

A large specific area of nanoporous glasses (up to 100 m2/cm3 and higher) provides a uniform spatial distribution of particles within porous sample, so that the particles immobilized on the matrix surface may be properly investigated. Amongst others the optical properties are of special interest because the optical sensors of the gas/atmosphere components have demonstrated its good performances and are widely used in the chemical analytical practice [4,5]. Particularly, spectral methods have proved its applicability and effectiveness here. In principal, these methods allow to detect optical signals produced even by single molecules [15].

First of all, the nanoporous glass is not anyhow a passive substrate for the adsorbed substance – on the contrary, the inherent matrix properties exert a critical impact on the composition properties on the whole. In contrast to the molecules being dissolved in the liquid solution, the same molecules being immobilized within highly developed nanoporous matrix behave substantially in different way. This situation causes certain peculiarities in spectral characteristics of immobilized molecules that are not event in case of molecules in solutions. Investigations of this matter might open good application prospects of the obtained results for developments of the optical chemical sensors including CO sensors.

As a complexing agent, the same Pd2+ ion was selected with taking into consideration the following ideas. This ion is able to form the number of intermediate carbonyl-halogenide coordination compounds and therefore to produce a gradual color transition of the composition samples. The Pd2+ ion has an electronic configuration d8, thus it might form the donor-acceptor bound with CO electron pairs.

2 Experimental

2.1 Sample preparation

Nanoporous samples were produced using the stepwise thermal and chemical treatment following a well-known Vycor process [9, 10]. Original solid glass was a liquated sodium-borate-silicate glass (sodium oxide – 6.8 mass. %, boron oxide – 26.7 %, silicon oxide – 66.0 %, the rest – other components) with phase separation. After having been treated in an aqueous solution of hydrochloric acid *HCl* (3N), a sodium-borate phase was leached out, thereafter the samples were annealed in air at +550 °C. Such procedures resulted in a through-out open porous structure within a substantially silica framework.

Next a procedure of the porous matrix impregnation followed in solution of the selected palladium salt. The procedure lasted until sorption equilibrium having been achieved. The salt molecules were forced to penetrate into nanoporous glass from the acetone, ethanol and dimethyl sulfoxide

solutions. The samples were then extracted from the initial solutions and dried in air at +50 °C in order to evaporate the solvent out of the matrix.

By means of described treatment the salt molecules have been fixed onto pores surface.

The samples were the plates of various thicknesses within 0.2 – 1.0 mm.

2.2 Spectral measurements

The absorption spectra of the samples and impregnating solutions were run on the spectrophotometers U-3200 (Hitachi, Japan), Lambda19 (Perkin-Elmer, USA) and СФ-26 (LOMO, Russia) within the wave range 300 – 800 nm.

During measurements, the samples were put into optically transparent cell being blown-through with CO containing air. Before taking readings, the samples were kept in the analysed gas mixture for sufficient time (10 min or longer). The reference spectral characteristics relate to the samples having been kept in normal laboratory atmosphere (relative humidity of 50 % at 20 ^{o}C).

3 Results and discussion

3.1 Studies of original nanoporous matrix

Small-angle X-ray techniques as well as a routine processing the adsorption isotherms in the area of capillary condensation have revealed the pore size distribution in vicinity of 8 – 9 nm [14], the porosity being of 28% of the total sample volume. Specific area, calculated from a specific bending point on the isotherm curve, turned out to be about 105 m^2/cm^3.

The studied porous glass exhibits an absorption isotherm of IV type (A-subtype) according to the Gregg's classification [16]. This fact witnesses the long-capillary pore structure. The resulted porous structure was permeable for gases and liquids, and the samples were practically transparent in the visible wave range.

3.2 Spectral studies of immobilized Palladium(II) chloride

In Figure1 (curve 1) an absorption spectrum of the initial impregnating medium is displayed. This medium being the acidulous water solution of $PdCl_2$. The spectrum shape is typical for the planar tetra-coordinated complexes of the Platinum row metals. Specifically, an absorption band at 350 nm may be assigned to transition connected with a charge transfer from ligand to metal in the dimer particles $[Pd_2Cl_6]^{2-}$ [17,18]. A broader band at 456 nm most probably belongs to the *d-d* transitions within Palladium ions. The curve 2 in Figure1 presents the spectrum of nanoporous sample containing the immobilized Palladium (II) chloride complexes. Lowering of the adsorption maxima may be caused by the restriction of the vibrational degree of freedom of particles immobilized on the pores surface.

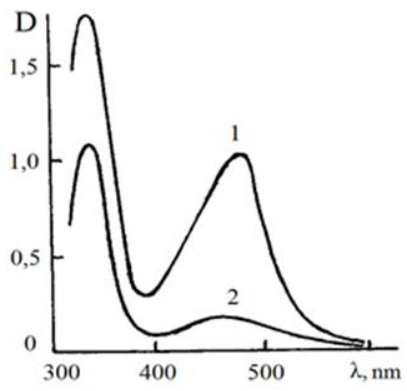


Figure 1: Absorption Spectra of the *$PdCl_2$* acidulous water solution (1 mass %)-(1); and nano-prous sample processed in this solution- (2)

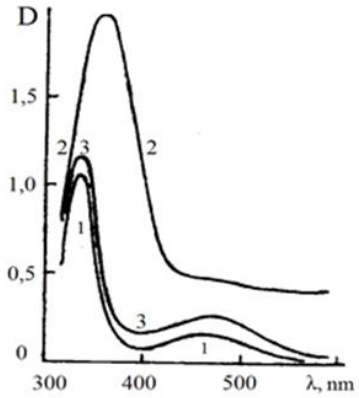


Figure 2: Absorption Spectra of the processed nanporous sample: in the initial atmosphere- (1); in the CO atmosphere (70 vol %)-(2); in the initial atmosphere again- (3).

In Figure2. a spectral response of the nanoporous sample on the carbon monoxide in atmosphere is presented. As one can see, the response was significant but rather slowed down. Air blowing the optical cell with clean air returned the spectrum to the initial shape, however, the sample spectrum recovered not entirely. Nevertheless the very fact of spectral recovery witnesses that metallic palladium reduction does not take place in this case. Most probably the intermediate complexes of the Palladium carbonyl-halogenides are formed on the dry pores surface. The most stable among them are the monomeric complexes *$[PdCOCl_3]^-$*, or dimer ones *$[Pd_2(CO)_2Cl_6]^{2-}$*. In general, the process may be written in such a way:

$$[PdCl_4]^{2-}{}_{surf.} \xrightarrow{+CO} [Pd(CO)Cl_4]^{2-} \xrightarrow{-Cl^-} \rightarrow$$

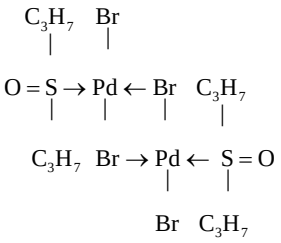
$$\rightarrow [Pd(CO)C1_3]^- \xrightarrow{-Cl^-} \rightarrow [Pd\,(CO)C1_2].$$

In essence, this reaction proceeds as an associative mechanism of the S_N2_{lim} type with formation of the rectangular pyramid. For avoiding the irreversible signal component, the glass surface should be chlorinated, however, it helped weakly.

In addition, the CO molecules could also take a bridge position between the adsorbed monomeric complexes. Some of them could remain fixed in this position on the pore surface providing an observed irreversible component of the spectral response. If so, cycling the CO attacks on the sample might improve the situation.

3.3 Spectral studies of immobilized Palladium (II) complexes

The encouraging results have been obtained with use of the more complicated compound such as Pd2-di-μ-dibromo-bis-(di-n-propyl sulfoxide):

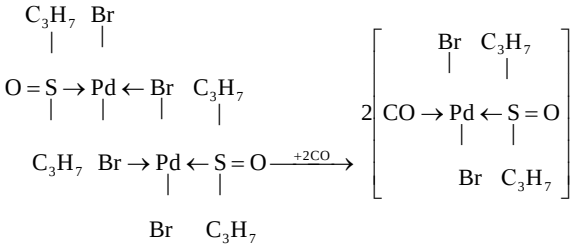


Noteworthy is the complex being initially a bridged dimer [19].

Spectra of the complex solutions in different solvents have been run. As an optimum solvent, dimethyl sulfoxide was selected because it's being homologous to dipropyl sulsoxide. Spectrum of

this solution is displayed in Figure3 (curve 1). Compared with Palladium (II) chloride, spectrum in Figure3 demonstrates the same two bands revealing presence of the dimeric molecules and electron transitions within Palladium(II) ions. However, the latter band is less expressed and being shifted to the shorter wave lengths.

These dimeric complexes are known to dissociate down to monomeric ones under action of the carbon monoxide molecules (see the reaction below) [20]. In Figure3 (curve 2) a spectrum of the bubbled solution is presented, one can see the lowering the dimer absorption band. Indeed, after barbotage procedure the solution gets clearer, and the precipitation process takes place. Surely, the sediment was thoroughly filtered off.



Taking into consideration such behavior of the system, we have used procedure of the porous samples impregnation in solutions of both non-bubbled and previously bubbled by the gaseous carbon monoxide. In Figure 4 the corresponding spectra are presented.

The nano-porous samples impregnated with that residual solution and then properly dried turned out to be potentially more suitable as the sensitive elements of the CO detectors. The more as the spectral response under the same conditions was almost entirely reversible. The surface reaction in this case does not proceed due to irreversible reduction of the Palladium ion by the CO molecules. The reaction mechanism includes only a bridging process accompanied by configuration changes in coordinative sphere of immobilized ion. Therefore optical response is exceptionally produced by CO molecules, and it provides very high selectivity of the sensor signal.

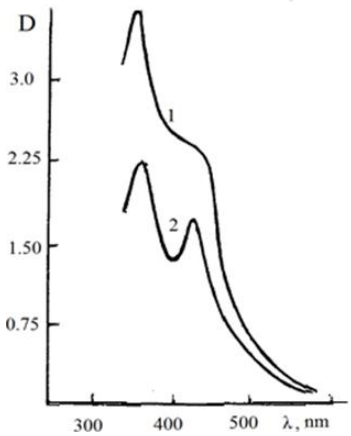


Figure 3: Absorption Spectra of the dimethyl sulfoxide solution of the Palladium (II) complex: 1-before barbotage; 2-after barbotage.

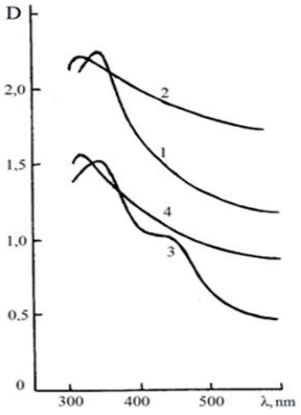


Figure 4: Absorption Spectra of the nano-porous samples processed in the dimethyl sulfoxide solution of the Palladium (II) complex: 1-sample from non-bubbled initial solution, normal atmosphere; 2-the same in the 70% CO atmosphere; 3 and 4 – the same from bibbled solution, respectively.

So, one can see that the CO concentration change by 70% produces the optical density change by about 0.5 (i.e. by 50%). Using the modern precise optoelectronic facilities, a high sensitivity to the CO appearance in atmosphere may be provided as well. Much better results might be achieved with use of a double wave lengths detection scheme (at 400 and 600 nm).

However, a time constant of the prototype sensors is rather high – about 5 min.

REFERENCES

[1] Dräger Werk AG. Product Catalog. Detector Tube Handbook, 2010.

[2] Thompson C. V., Goedert M. G. Field-Portable Instrumentation for Gas and Vapor Samples. In: Encyclopedia of Analytical Chemistry / Ed. by Meyers R. A., John Wiley and Sons, Inc. Vol. 14, Field-portable Instrumentation, 2009.

[3] Opekar F., Štulík K. Electrochemical Gas Sensors. Ibid. Vol. 9, Electroanalytical Methods, 2009.

[4] Lucena R. Infrared Sensors. Ibid. Vol. 20, Infrared Spectroscopy, 2010.

[5] Saltzman R. S. Ultraviolet/Visible Spectroscopy in Process Analyses. Ibid. Vol. 33, Process Instrumental Methods, 2011.

[6] Holleman A. F., Wiberg E. Inorganic Chemistry. Academic Press. San Diego, 2001.

[7] Livingstone S.A. The Chemistry of Ruthenium, Rhodium, Palladium, Osmium, Iridium and Platinum, Oxford: Pergamon Press. 1973.

[8] T. H. Allen, W. S. Root. Colorimetric Determination of Carbon Monoxide in Air by an improved Palladium Chloride Method. // Journ. Biol. Chem. 1955. V. 216, No.1, pp. 309–317.

[9] Yanowski F., Heyer W. Poröse Gläser. Herstellung, Eigenschaften, Anwendung. I Auflage. Leipzig: VEB Deutscher Verlag fűr Grundstoffindustrie. 1981. 276 S.

[10] Enke D., Janowski F., Schwieger W. Porous glasses in the 21st century − a short review // Microporous and Mesoporous Materials. 2003. V. 60. N 1−3. P. 19−30.

[11] Macedo P. B., Litovits T. A. Method of precipitation of dopants in a porous silicate glass. – Patent USA No. 4110096, publ. 29.08.78.

[12] Novikov A.F. Nanoporous silica glass sensibilisation in respect to the gas components detection. // Optica Applicata, 2008, V. XXXVIII, No.1, 65-69.

[13] Elmer T. H. Porous and Reconstructed Glasses. In: Schneider S. J. (ed.). Engineered Materials Handbook, Vol. 4: Ceramics and Glasses. Materials Park, OH: ASM International. 1991, pp. 427–432.

[14] Novikov A.F. Characterization of the inner structure and surface of nanoporous sodium-borate-silicate glasses. // Optica Applicata, 2005. V. XXXV, No.4, pp.702-708.

[15] Basche T. [ed.], Single-Molecule Optical Detection, Imaging and Spectroscopy. VCH Publ. 1997, 250 p.

[16] Gregg S., Sing C. Adsorption, Surface Area & Porosity. L. – N.Y.: Academic Press. 1967.

[17] Lever A.B.P. Inorganic Electronic Spectroscopy. 2nd ed. Part 2, section 6.2.8. Elsevier, ser. Studies in Physical and Theoretical Chemistry 33. 1985.

[18] Rush R.M., Martin D.S., Jr., Le Grand R.G. Electronic spectra of the Pd complexes. // Inorg. Chem. 1975. V.14, № 10, pp. 2543–2550.

[19] Maitlis P. M. The Organic Chemistry of Palladium. N.Y.: Academic Press. 1971.

TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

An Enriched Ciphering Method to Evaluate Performance of EEA2-algorithm for LTE Security

Gautam Siwach[1], Amir Esmailpour[2], and Ahmad Sharifinejad[3]
[1,2]Department of Electrical and Computer Engineering,
University of New Haven, West Haven, CT, USA
[3]IEEE member, Oslo, Norway
gsiwach@gmail.com, aesmailpour@newhaven.edu, a_sharifinejad@yahoo.no)

ABSTRACT

In order to address vulnerabilities in LTE security and the objective to strengthen the implementation of encryption and decryption algorithms in LTE, we proposed a method for integration of an algorithm with a dynamic matrix generation scheme of 16*16 elements during AES implementation. Implementing such algorithm eradicates the vulnerability that causes a hacker to access plain text. The vulnerability arises within the implementation of AES.

This research explores Enriched Ciphering algorithm for number of scenarios based on function calls and time metrics that increases 3.9 times when there is a change in size of the plain text. We evaluated the proposed algorithm referred to as Enriched Ciphering Algorithm based on various use cases, including compression and decompression of cipher text that provides an extra layer of complexity to strengthen the security, and provides performance improvements of 13.9 percent when executing the Enriched Ciphering algorithm with big data size.

Keywords- AES, LTE, EEA1, EEA2, PDU.

1 Introduction

In this study we introduce a new algorithm to address some security vulnerabilities in the Long-Term Evolution (LTE) technology caused by compromised encryption/decryption process. Our new proposed algorithm so called “Enriched Ciphering” is optimized for performance evaluation. This approach offers a solution to potential vulnerability in original EEA1 and EEA2 algorithms. The vulnerability of these algorithms is due to having plain text in the remaining part of cipher text within one PDU, which is exposed to intruders.

Advanced Encryption Standard (AES) can be performed on blocks of 128 bits of data according to the National Institute of Standards and Technology (NIST) standards. Once the data is encrypted by AES, and a cipher key generated then XOR operator is applied on the plain text part of PDU in order to construct a Ciphered Text (CT). In order to apply the described procedure to our payload data, the Plain Text (PT) is partitioned into blocks of 128-bit data as shown in equation (1). The final block need not be 128 bits.

$$PT = PT[1] + PT[2] + \ldots + PT[n] \quad (1)$$

Encryption of payload happens by using AES Counter (AES-CTR), in order to strengthen the security. Each PT block is XOR’ed with a block of key stream to generate the (CT). The AES counter encryption of each Plain text block results in 128-bits of key stream because it is partitioned into blocks of 128-

DOI: 10.14738/tnc.32.1092
Publication Date: 25th April, 2015
URL: http://dx.doi.org/10.14738/tnc.32.1092

bit data. The most significant 64 bits of the counter block PT are initialized as seen before, followed by 64 bits that are all 0. The AES function performs AES encryption under the control of the confidentiality key. (Confidentiality Key refers to the private key). In order to make sure that even final block is partitioned to the appropriate length of 128-bit data, we use the TRUNC function; it truncates the last output of the AES encryption operation to the same length as the final PT block, returning the most significant bits.

The decryption operation is similar to the encryption, and the AES-CTR uses the same AES encryption operation since AES is a standard algorithm and it does not change. So of course the same AES applies both ways for both encryption and decryption, we are not making any modification to AES algorithm we are applying AES to LTE stream to make it more secure.

We execute the new edition EEA2 algorithm that includes a matrix as an enhancement to the ciphering process called enriched ciphering. The purpose of this study is to present a secure implementation for an existing vulnerability in the LTE encryption technique and to conduct a performance evaluation in order to provide a benchmark for further improvements.

2 Background

AES algorithm implementation has a potential vulnerability. In [10] “LTE Security Potential Vulnerability and Algorithm Enhancements”, we presented the vulnerability and provided the design as a solution for it through the enriched ciphering algorithm. The breakdown for explaining the vulnerability is let us assume that we have only acquired the entire cipher text by using brute force attack, and we don't have any information on either the plain text or the key. Evidently we need more information on the original plain text or the key to be able to go any further in deciphering the content. Following this let us assume that we now have captured the plain text and cipher text within a PDU, so we are missing the key to be able to recover the entire plain text. Finally assume that if we have retrieved 128 bits of plain text then we will have the corresponding cipher text of 128 bits, and as well as the remaining cipher text.

We have described the complete vulnerability and proposed the design of a solution to address it, we implemented the proposed solution including the main encryption mechanism key generation and distribution in Matlab simulation software. The main component of the solution are presented in the Enriched-ciphering algorithm, the AES algorithm could rebuild the key when we know plain text and corresponding cipher text. The input for AES is fixed within one PDU, so that the outputs from AES are the same length as input, which will be used to XOR with the plain text. The vulnerability was that once we get the output as the cipher text per PDU and complete cipher text, then we could retrieve the plain text by simply using XOR operation between the complete cipher text and the output. As per our assumption, we get 128 bits of plain text and the corresponding cipher text and the remaining of the cipher text. We XOR the plain text and cipher text; we get the output from AES. Upon using this output to XOR with the cipher text the remaining plain text could be regenerated.

3 Literature review

The need for enhancement in authentication services is motivated due to a fact that even Password Authentication Protocol (PAP) and Lightweight Extensible Authentication Protocol (LEAP) are vulnerable as investigated in [1] by Lianfen Huang, et.al. They experimented on authentication protocols in LTE Environments and summarize authentication service as the most important services in LTE networks in “Performance of Authentication Protocols in LTE Environments.”

In [2] Bin Liu et.al investigates the implementation issues in "Parallel AES encryption engines for many core processor arrays, by exploring different granularities of data-level and task level parallelism, based on the core systems largest with 107 to 137 cores and depicting several AES implementations on a fine-grained many-core system. They concluded that design on a fine-grained many core system software platforms achieves energy efficiencies approximately 2.9-18.1 times higher as compared to other software platforms. By this theory we expect that the designed solution of adding a matrix to the AES encryption can be evaluated for performance improvement.

There are several researches for LTE, AES, and also the integration of several other platforms like VLSI, Graphical user interface and many core cross platforms those served as a motivation for this research In [3] Naga M. Kosaraju et al illustrates the AES Algorithm architecture during their research for "A high- Performance VLSI Architecture for AES Algorithm." They presented architecture by expanding the secret key used to generate subkeys in VLSI implementation of AES algorithm. They implemented a prototype chip using 0.35 µ CMOS technology resulting in a throughput of 232Mbps for iterative architecture and 1.83Gbps for pipelining architecture. They explore the architecture for key scheduling Unit and are yet another instance to prove the integration and the combination logic.

In addition, AES takes a lot of time for encrypting which has been a major concern of all the researchers like [4] Fei Shao et al during "AES Encryption Algorithm based on High Performance Computing of GPU." They have proposed different approaches in order to expedite the encryption process. Simultaneously [5] Keisuke Iwai et al in "AES encryption implementation on CUDA GPU and its analysis" define granularity and memory allocation as major contributors to effective processing in AES encryption on GPU, thereby supporting the former content. Keeping this as context, we provide a real run time graphical illustration in the results section of this paper. We use different figures for illustrating the analysis of the performance and for future work.

There had been several investigations to ensure the privacy in LTE. One of them is by [6] Khodor Hamandi et al in "Privacy Enhanced and Computationally Efficient HSK-AKA LTE Scheme," in which they explain the HSK-AKA procedure and the key management system from HSS to UE by using a periodic temporary identifier. They gave an illustration of how the process flow occurs between HSS and UE and presented a modified LTE Authentication and Key Agreement (AKA) scheme,HSK-AKA.

As iterated by [7] Mehran Mozzaffari-Kermani et al in "Concurrent structure-Independent Fault Detection Schemes for the Advanced Encryption Standard" AES has been lately accepted as the symmetric cryptography standard for confidential data transmission. However, we still strive for reliable AES architecture. Their investigation concludes with the implementation of a structure independent scheme in order to reach the error coverage of approximate 100%.

The need for analyzing compression techniques along with encryption is an important consideration and is also stated in "Securing multimedia content using Joint compression and encryption" authored by [8] A. Pandae at al. They state that the algorithmic parameterization and hardware architectures are useful to ensure secure transmission of multimedia data in resource-constrained environments such as wireless video surveillance networks. The "Enriched Ciphering algorithm" is executed along with compression and is analyzed in the context of time complexity as well.

As explored by [9] Krzysztof Jankowski et.al in "Packed AES-GCM Algorithm suitable for AES/PCLMULQDQ instructions" about the performance evaluation of authenticated Encryption Algorithms that AES block cipher working in Counter Mode (CTR), characterized with key sizes 128, 192, and 256 is used to XOR the Encrypted Counter with the plain text. They craft a gateway to

further investigate the approach for parallel encryption and increment. Moreover, we support their investigation and plan to advance this research for parallel encryption.

4 The enriched ciphering method design and implementation

The design of the Enriched Ciphering algorithm is to ensure security during the process. In [15] we also presented a security structure of a ciphering algorithm that will build a matrix on the concept of generating random numbers based on permutation; the matrix will be incorporated in the encryption process by including it in the final stage of encryption algorithm in order to get the enriched cipher text. A matrix block of size 16*16 is used. Within each block of 256 elements, there are 8 binary bits, which is equal to 2 hex digits. The 256 elements are different and make the enriched cipher text. This process will enhance encryption by providing a more secure cipher text, which is not as vulnerable as the original encryption process, hence labeled: “Enriched cipher text”.

For this Enriched Ciphering algorithm, the block used depends on the cipher text stage1, and the cipher text stage1 is the plain text XOR’ed with the key stream that has dynamic elements generated by matrix. It requires a significant amount of data to break this Enriched Ciphering algorithm. In addition, the key stream will be changed per PDU. Therefore, this process will not disclose enough data for the intruder to be able to crack the Enriched Ciphering algorithm easily. However, for the same instance, considering 128 bits of plaintext and corresponding cipher text, the derivation of the key stream will be far easier. This indicates a classical trade-off scenario between complexity and space used, the higher the capacity the larger space is occupied by the data stream as data is divided per PDU and the key is generated for every 128 bit of plain text.

4.1 Key generation mechanism

In this section we describe the process of generating the key. This is to generate a random number of keys for 256 elements of the matrix block. We generate the reverse matrix based on the original matrix and a new empty matrix for each block of the original matrix and obtain the value. The first digit is the column number for reverse matrix, and the second digit is the row number for the reverse matrix. Once we get a block in reverse matrix, we fill it by a combination of the original column number and row number, by repeating this process for each block, the reverse matrix will completely be rebuilt.

$$\text{Key=key (randperm (numel(key)))} \tag{2}$$

4.2 Distribution of the key

The key is 128 times larger than the normal key that has the pain text of 128*128 bits size, so it is not easy to share the key with confidentiality. Therefore we just share it before using EEA2 that happens before NAS and AS signaling. UE generates the matrix and just encrypts it by AES and sends it to the MME before NAS and to the eNodeB before AS signaling. Since both MME and eNodeB can calculate the reverse matrix, the UE can use EEA2 between MME and eNodeB. We can choose the KASME for AES to MME, and for AES to eNodeB. We can choose the KeNB because these two keys were already shared before so they are used for mutual authentication.

4.3 Implementation

The implementation is based on the design and the key generation mechanism and the distribution of key concepts, combined in the Enriched Ciphering algorithm. The code uses the permutation concept to generate random numbers for elements of matrix block, it includes conditional

statements and logic gates operations as well. Moreover, the matrix is able to provide more secure operations in terms of encryption complexity.

Matrix is dynamically generated and processes 128 bits round of plain text. In an example the plain text is passed as arguments through the matrix and is divided into 128 bits of text as one round that is transformed as per hexadecimal notion by Enriched Ciphering algorithm as step I.

The Enriched Ciphering algorithm provides a solution for possible vulnerability within the encryption process, which involves adding a matrix block to the encryption process of the cipher text results obtained upon executing the Enriched Ciphering algorithm in different scenarios. The encryption process uses a dynamic set of numbers in the form of a matrix, which is used with the cipher text to obtain an enriched block of ciphered data.

In step II the plain text is processed to cipher text. This cipher text is encrypted in step III by a matrix. 'OutCT' or Output cipher text of 128 bits shows the text after encryption by the matrix. backCT shows the decrypted text and the loop follows till the end and whole plain text is retrieved upon reducing it to its normal form. We Evaluate and analyze the procedure, design and algorithm. The performance evaluation is based on different performance evaluation metrics such as time, call, self-time and function call. 'Call' refers to calling the predefined function at a particular instance in the program. The self-time and time are the sub parts of the total time taken to execute the Enriched Ciphering algorithm. One of the sample test run of the Enriched Ciphering algorithm is depicted in this paper for performance evaluation. It is an instance run and is utilized to evaluate the performance of the Enriched Ciphering algorithm.

Enhancements occur upon adding compression and decompression functions to the existing Ciphering algorithm. These functions are included as a part of Enriched Ciphering algorithm in an effort to improve the performance even when the text is larger in size so as to make space for faster, efficient ways to build strength. The increase in the execution time of the Enriched Ciphering algorithm is associated with the number of calls for functions with increased size of text, including total time and self-time of processes.

Since the key elements of the matrix keep on changing and are dynamic in order to provide a unique security level, so the run time can't be generalized for all Enriched Ciphering algorithm executions. We provide the precise time of a particular run based on different scenarios in the cases below. Based on the single instance execution and associated outcomes, we comment on the performance improvements and analyze the complexities of the Enriched Ciphering algorithm.

The proposed modifications in the algorithm is presented below as a Pseudo code:

1: Define a cipher key

2: Assign number of elements of the matrix

3: Apply the permutation logic to generate the key elements

4: Input plain text

5: Compress the plain text

6: Disintegrate the input text into 128 bits round of text until the end

7: Input data type

8: BITXOR the data

9: Reshape the matrix

10: Retrieve the Cipher text

11: Encrypt the data

12: Use key to decrypt the encrypted data

13: Convert the cipher text into plain text

14: Use rotation and BITXOR

15: Decompress the data

16: Retrieve the plain text

5 Results

In this section we consider several test-cases to evaluate performance of the enriched ciphering method. The different cases considered in process of evaluating the performance starts with executing a round of plain text that has length of 128-bit. Self-time is the time spent in a function excluding the time spent in its child functions. Self-time also includes overhead resulting from the process of profiling. Total time is the self-time and time taken to run child functions as well. Call refers to the number of times function is called. dec2Hex, hex2dec, iscellstr are the internal functions in Enriched Ciphering algorithm the significance of these functions is that they are required to change key values when making matrix dynamic.

Case I: The instance described in Table 1 shows the numbers of function name and call. The call decides for the time. The functions are used while executing the Enriched Ciphering algorithm. They have a direct effect on run time of the Enriched Ciphering algorithm. In Figure 1, we show the self-time and real time on the axis and with a scale of .002 seconds. The Enriched Ciphering algorithm executes completely in .017s, sufficient time for the execution by the proposed solution.

Table I: Time slot details upon running the Enriched Ciphering algorithm

Function Name	Call	Total Time (Seconds)	Self-Time (Seconds)
Enriched Ciphering Algorithm	1	.017	.011
dec2hex	1	.004	.004
hex2dec	1	0.003	.003
iscellstr	1	0	.000

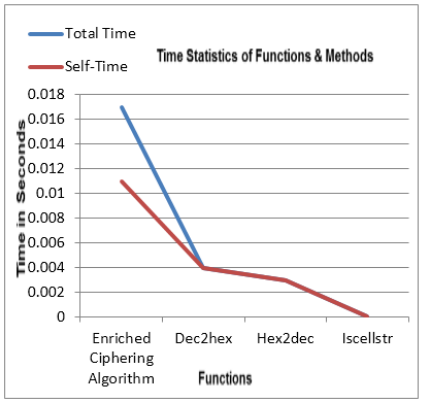


Figure 1: Illustration of total time and self-time in unit of seconds for a round of plain text, and different functions in the Enriched Ciphering algorithm

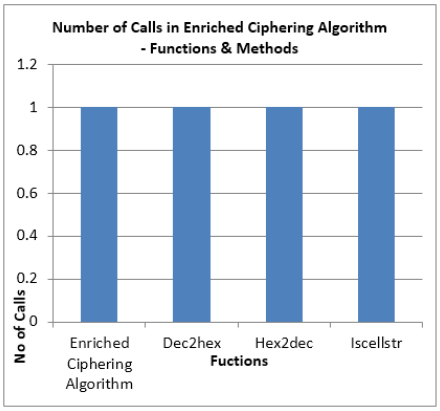


Figure 2: Illustration of number of calls for the functions associated to execution of Enriched Ciphering Algorithm

Case 2: In another case, the size of the string is increased and it is noticed that the changes occur in the execution of the Enriched Ciphering algorithm; Figure 3 shows us an increase in the number of calls upon increasing the length of the text size, when we notice the changes in terms of time complexity. Hence, we observe that there are more associated function and method calls for large size of data.

Figure 4 shows the increase in total execution time as per the increase in length of the string; although the changes are not constant; however the execution time is proportional to the length of the string which refers to the case when we execute the Enriched Ciphering Algorithm keeping the size of the string larger than 128 bits rounds of text as compared to case 1.

After execution of the Enriched Ciphering Algorithm the results show us an increase in the execution time upon increasing the length of the text; hence the execution time of Enriched Ciphering algorithm is proportional to the number of rounds per bits of text. Although the number of functions and methods called during the execution of Enriched Ciphering Algorithm are almost the same in contrast to the number of functions called by case 1.

The number of calls of the associated functions in Case2 is more than that in case 1 because the size of the string is increased from one round of text which has length of 128 bits then to the two round of text which is of length 256 bits approximately. The total time of execution of Enriched Ciphering Algorithm is more in case 2 as compared to the time of execution of Enriched Algorithm in Case 1. It is .084 seconds in this case and .017 seconds in the former. Through these cases we determine and evaluate the performance of the Enriched Ciphering Algorithm with different cases and present the potential areas for performance improvement.

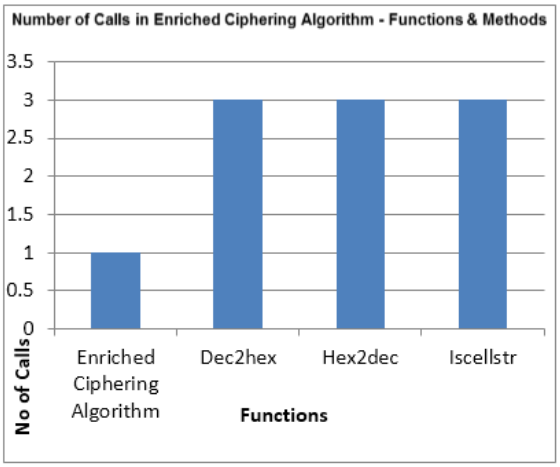


Figure: 3 Illustration of number of calls for the functions associated to execution of Enriched Ciphering algorithm with a larger string size.

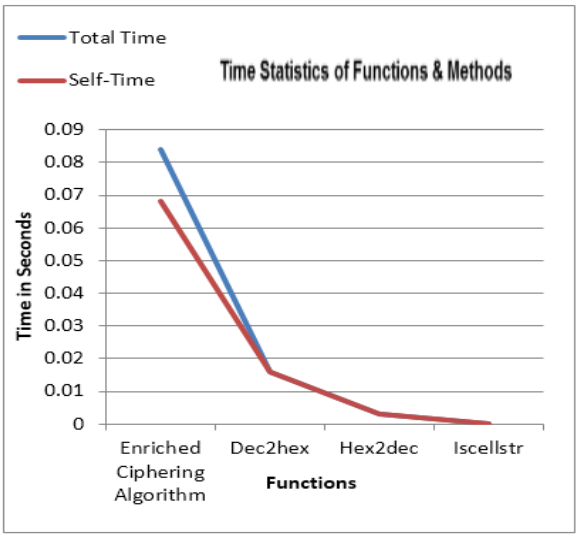


Figure 4: illustration of total time and self-time in unit of seconds for different functions in the Enriched Ciphering algorithm for larger size of text

Table 2: Time slot details upon running the Enriched Ciphering algorithm and increasing the length of the plain text; Size Increase in the length of Plain

Function Name	Call	Total Time (Seconds)	Self-Time (Seconds)
Enriched Ciphering Algorithm	1	.084	.068
Dec2hex	3	.016	.016
Hex2dec	3	0.003	.003
Iscellstr	3	0	.000

Case 3: A further change of the enhancement in the proposed solution relates to large text files that can be processed in a secure and fast manner. In order to address memory issues we suggest the use of compression technique in the Enriched Ciphering algorithm and analyze the time domain at the runtime execution.

Figure 5 illustrates the number of calls associated to the functions in Enriched Ciphering algorithm per round of text when zip function is used. Create Archive and get archive are the methods and functions associated to Enriched Ciphering algorithm use to store and call values while execution of Enriched Ciphering algorithm.

Table 3: Function details upon integration of compression technique with the Enriched Ciphering algorithm, and analysis of statistics with path of associated functions a method.

Function Name	Call	Total Time (Seconds)	Self-Time(Seconds)
Enriched Ciphering Algorithm	1	.147	.081
Zip	3	.066	.000
Create	3	.066	.000
Get	3	.049	.000
Unique	6	.017	.000
Fileparts	9	.017	.017
Add	3	.017	.000
(Java Method)	15	.017	.017
Legacy	6	.017	.017
Fullfile	3	.016	.016

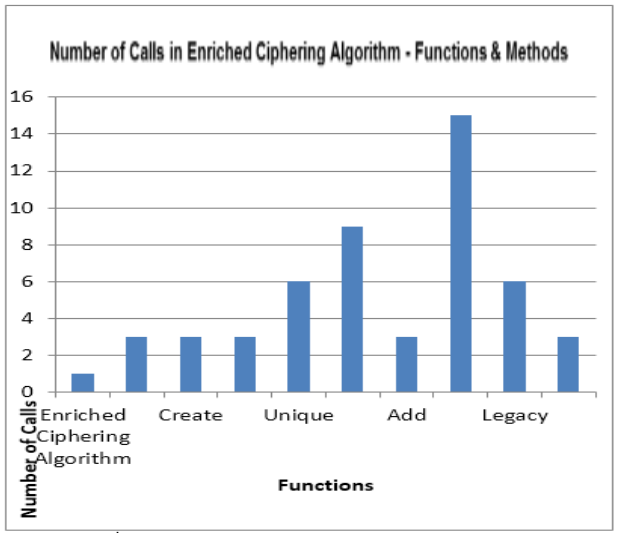


Figure 5: Illustration of number of calls when Enriched Ciphering algorithm execution involves zip function per round of text.

This case is based on an approach to obtain the overhead involved in executing the zip function at the initial level of a round of text. With this we investigate the use of encryption technique in an Enriched Ciphering algorithm when the text is small. We notice the increase in execution time and number of calls for functions.

Case 4: Compression technique stands for the memory concerns for the users. Now, finally there is an increase in the runtime of an Enriched Ciphering algorithm when using compression technique in an effort to address storage and memory issues. Here, in Case 4, our exploration of the Enriched Ciphering algorithm begins with increase in length of the string. At Table 4, the functions, and the calls appear to be more also illustrated in Figure 7. But there is a decrease in execution time as per Figure 8. Hence, we determine that compression technique yields efficient results after runtime execution when a larger size of text is involved in the process and it addresses memory concern as well.

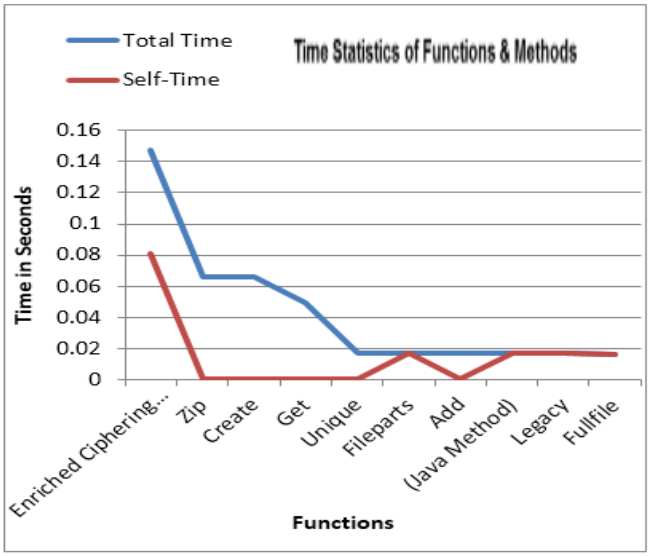


Figure 6: Illustration of total time and self-time in unit of seconds for different functions in the Enriched Ciphering algorithm when compression is involved for an initial round of text.

Table 4: Compressing and increasing the length of the plain text, and evaluating the time statistics with path of associated functions and methods

Function Name	Call	Total Time (Seconds)	Self-Time (Seconds)
Enriched Ciphering Algorithm	1	0.127	0.063
Zip	1	0.048	0
Create	1	0.048	0
Get	1	0.032	0.015
Unique	2	0.016	0.016
Dec2hex	1	0.016	0.016
Add	1	0.016	0
Get	3	0.016	0.016

Case 5: - In this scenario, we include the decompression step in order to calculate the runtime execution of the Enriched Ciphering algorithm. We integrate the unzip function to the scenario executed in Case 4 of this section, in the Enriched Ciphering algorithm and execute it.

Upon execution, we get the runtime details as mentioned in Table 5. This is the case of compression and decompression technique on a round of text where new functions also take part upon in the processing of Data upon execution of Enriched Ciphering Algorithm. Since there are few new processes involved in the Enriched Ciphering algorithm because of the changes done in the length of string and also due to the new functions and methods involved in order to evaluate the performance of the ECA.

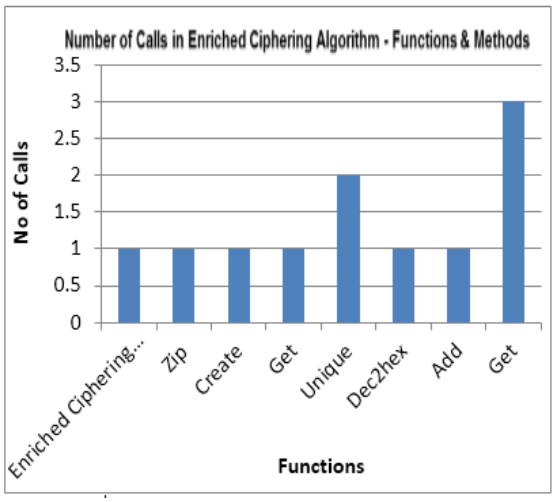


Figure 7: Illustration of number of calls when Enriched Ciphering algorithm execution involves zip function for a larger size of text

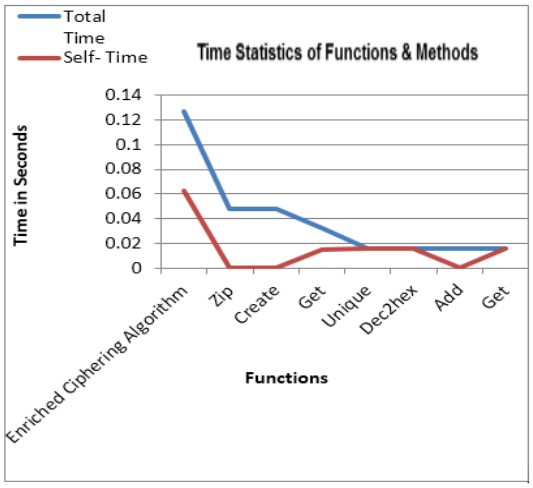


Figure 8: Illustration of total time and self-time in unit of seconds for different functions in the Enriched Ciphering algorithm when compression is involved for a larger size of text

Figure 9 shows the number of calls for the functions and methods involved in the compression and decompression technique upon integration within the Enriched Ciphering Algorithm, and when the Enriched Ciphering algorithm is executed then the indicated calls shows the execution statistics of the associated child functions as well. These associated functions appear because of their association with the decompression stage in the Enriched Ciphering algorithm.

Create, Get and add are some of the functions and methods for the Archive entries of the files having data, whereas parse, Dec2hex, Num2Str indicates the parsing, processing, and conversion of different data types. The input string processing of the file takes place in splits of data carried by File parts. Zip and Unzip functions are the direct operation on the specific set of Data in order to save memory allocation and fine tune the performance on large set of Data.

Table 5: Decompressing a round of text and evaluating the time statistics with path of associated functions and methods

Function Name	Call	Total Time (Seconds)	Self-Time(Seconds)
Enriched Ciphering Algorithm	1	0.155	0.078
Zip	1	0.047	0.015
Create	1	0.032	0
Dec2hex	1	0.016	0.016
Get	1	0.016	0
Add	1	0.016	0
Convert	4	0.016	0.016
Num2str	1	0.016	0
Int2str	1	0.016	0.016
Mode	1	0.016	0
Get	3	0.016	0
Parse	1	0.015	0
Fileparts	4	0.015	0.015
Unzip	1	0.015	0

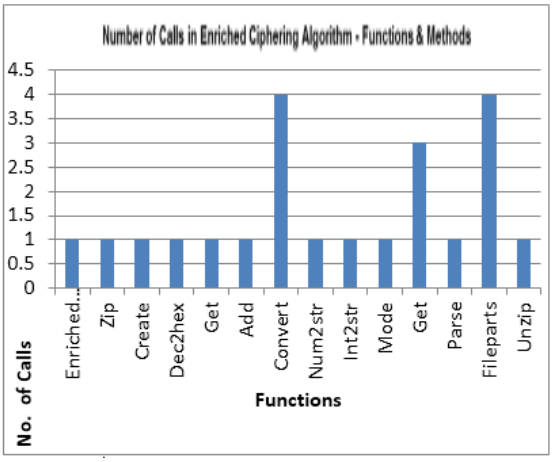


Figure 9: Illustration of number of calls with zip and unzip functions for a round of text in ECA.

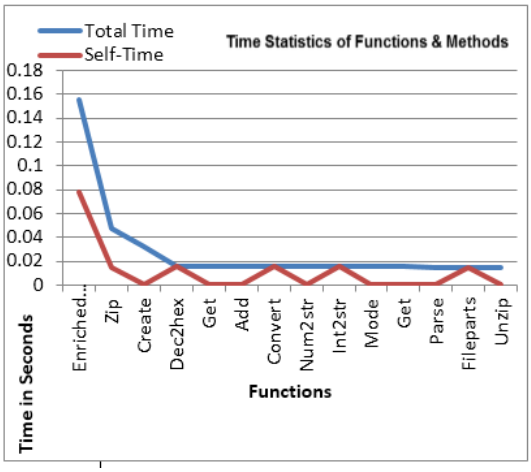


Figure 10: Illustration of total time and self-time in unit of seconds for different functions in the Enriched Ciphering algorithm when compression and decompression techniques are involved for an initial round of text.

Compression and decompression are mostly intended to save memory usage in a digital world. The time statistics described in Figure 10 and related to self-time and total time help us to analyze a slight increase in the curve of time upon including the decompression technique within the Enriched Ciphering algorithm.

Case 6: With larger text, the compression and decompression functions need to be emphasized more. We notice that there has been an increase in functions involved. In each function, the numbers of calls precede the function calls for smaller size of text. Table 6 gives the information of the number of calls, and Figure 12 provides a graphical view of the number of calls associated with the functions involved, and those called during runtime execution of the Enriched Ciphering algorithm. It is 0.32 seconds of time the Enriched Ciphering algorithm takes for complete execution (see Figure 11).

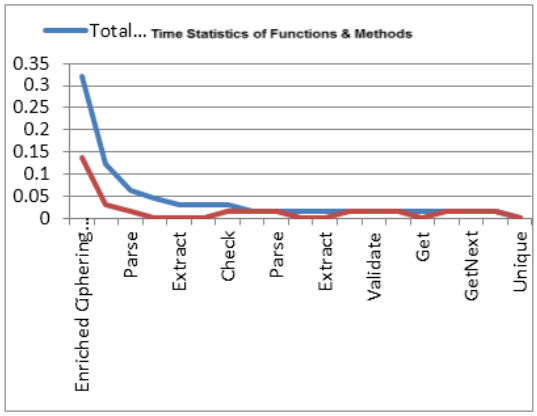


Figure 11: Illustration of total time and self-time in unit of seconds for different functions in the Enriched Ciphering algorithm when compression and decompression techniques are involved for larger set of Data and an initial round of text.

Table 6: Details of runtime execution of Enriched Ciphering algorithm upon integrating decompression with the larger text in order to evaluate the time statistics with path of associated functions and methods.

Function Name	Call	Total Time (Seconds)	Self-Time (Seconds)
Enriched Ciphering Algorithm	1	0.321	0.136
Unzip	3	0.123	0.03
Parse	3	0.062	0.016
Zip	3	0.046	0
Extract	3	0.031	0
Create	3	0.03	0
Check	3	0.03	0.015
Hex2dec	3	0.016	0.016
Parse	3	0.016	0.016
Add	3	0.016	0
Extract	3	0.016	0
(Java method)	9	0.016	0.016
Validate	3	0.016	0.016
Converte	3	0.016	0.016
Get	9	0.016	0
Isdir	3	0.016	0.016
GetNext	6	0.016	0.016
GetArc	3	0.015	0.015
Unique	6	0	0

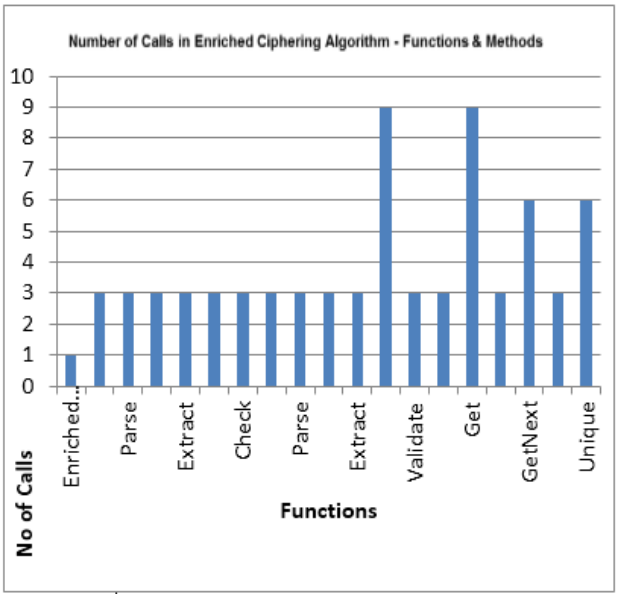


Figure 12: Illustration of number of calls when Enriched Ciphering algorithm execution involves zip and unzip functions for a larger set and per round of text.

6 Conclusion

We investigated and evaluated the Ciphering algorithm to add matrix block in order to strengthen LTE security. In this paper we present the results for performance evaluation based on different performance vectors. The results presented are real runtime execution and accurate. The simulation provides an estimated time increase of .067s being 3.9 times original of .017s execution time of Enriched Ciphering Algorithm when text size increases from 128 bits to 256 bits.

Moreover, we provide a real time listing of function calls during execution in the quest of modularizing the complete process to embed security at each phase. We also determine the increase in calls, and those are involved and appear into the system. Besides representing several other cases, we also provide the decrease of 13.60% upon including compression technique on large size of data sets to that of smaller size. We propose including compression and decompression for performance improvement to be a part of the process especially when we are dealing with large chunk of data sets so as to meet our memory allocation needs. We represent our research on integrating new techniques in the implementation and the need of improvement and strengthening LTE Security.

REFERENCES

[1]. "Lianfen Huang*, Ying Huang, Zhibin GAO, Jianan Lin, Xueyuan Jiang", Performance of Authentication Protocols in LTE Environments, 2009 International Conference on Computational Intelligence and Security, 293-297,978-0-7695-3931-7/09 © 2009 IEEE DOI 10.1109/CIS.2009.50.

[2]. "Bin Liu, IEEE, and Bevan M. Baas.", Parallel AES Encryption Engines for Many-Core Processor Arrays,"IEEE TRANSACTIONS ON COMPUTERS, VOL. 62, and NO. 3 MARCH 2013", 536-547, www.computer.org/publications/dlib.

[3]. "Naga M. Kosaraju, Murali Varanasi, Saraju P. Mohanty", A High-Performance VLSI Architecture for Advanced Encryption Standard (AES) Algorithm, Proceedings of the 19th International Conference on VLSI Design (VLSID'06) ,1-4,1063-9667/06 © 2006 IEEE.

[4]. "Fei Shao, Zinan Chang, Yi Zhang", AES Encryption Algorithm Based on the High Performance Computing of GPU, 2010 Second International Conference on Communication Software and Networks,588-590,978-0-7695-3961-4/10 © 2010 IEEE DOI 10.1109/ICCSN.2010.124.

[5]. "Keisuke Iwai and Takakazu Kurokawa, Naoki Nisikawa", AES encryption implementation on CUDA GPU and its analysis, 2010 First International Conference on Networking and Computing, 209-214,978-0-7695-4277-5/10 © 2010 IEEE DOI 10.1109/IC-NC.2010.49.

[6]. "Scheme,Khodor Hamandi Imad Sarji Ali Chehab Imad H. Elhajj Ayman Kayssi", Privacy Enhanced and Computationally Efficient HSK-AKA LTE, 2013 27th International Conference on Advanced Information Networking and Applications Workshops,929-934,978-0-7695-4952-1/13 © 2013 IEEE DOI 10.1109/WAINA.2013.133.

[7]. "Mehran Mozaffari-Kermani, and Arash Reyhani-Masoleh", Concurrent Structure-Independent Fault Detection Schemes for the Advanced Encryption Standard, "IEEE TRANSACTIONS ON COMPUTERS, VOL. 59, NO. 5, MAY 2010",608-622,0018-9340/10/ ©2010 IEEE.

[8]. "Amit Pande and Prasant Mohapatra", Securing Multimedia Content Using Joint Compression and Encryption ,2013 Oct- Dec Published by the IEEE Computer Society,50-61,1070-986X/13/ c 2013 IEEE.

[9]. "Krzysztof Jankowski and Pierre Laurent", Packed AES-GCM Algorithm Suitable for AES/PCLMULQDQ Instructions, "IEEE TRANSACTIONS ON COMPUTERS, VOL. 60, and NO. 1, JANUARY 2011", 135-138, 0018-9340/11/ 2011 IEEE Published by the IEEE Computer Society.

[10]. "Gautam Siwach, Amir Esmailpour ", LTE Security Potential Vulnerability and Algorithm Enhancements, 2014 IEEE Canadian Conference on Electrical and Computer Engineering, 1-7, IEEE CCECE May 2014.

TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

A Mobile Dual VoIP System for Enhancing Speech Quality and Intelligibility: Simulation and Test Bed

Francesco Beritelli and Corrado Rametta
Dipartimento di Ingegneria Elettrica Elettronica e Informatica, University of Catania, Italy
name.surname@dieei.unict.it

ABSTRACT

As it is well known in a 3G/4G network scenario, the quality of voice traffic over IP (VoIP) is greatly reduced due to the strong current limitations in terms of requirements regarding delay, jitter, packet loss rate and guaranteed bandwidth. The present work highlights the benefits in terms of improved intelligibility when making a duplication of VoIP packets through two wireless data accesses provided by different operators. In particular, the paper presents the architecture and the prototype of a dual stream approach to mobile VoIP applications (Dual VoIP) over HSPA access networks. Test results, obtained via simulations and a real-time implementation of Dual VoIP algorithm, demonstrate an average packet loss reduction up to 90% and an average improvement of speech quality up to 1 PESQ point. Furthermore, the paper highlights the significant reduction of the audio signal clipping at all levels: phoneme, word, sentence and conversation. Enhancement of the speech quality and intelligibility of the audio signal is a very important aspect in common best effort applications using VoIP as well as in particular conditions such as environmental wiretapping for forensic uses and/or private tactical communications in network-centric contexts where real time listening and intelligibility of the speech signal play a key role. The deep evaluation presented here has the aim of understanding the behavior of the proposed architecture under different application scenarios and drawing, at the same time, useful conclusions on the improvement of the Dual VoIP prototype.

Keywords: Mobile VoIP; Speech Quality Enhancement; Audio Clipping Removal; Packet Duplication; HSPA Networks.

1 Introduction

The rapid and continuous consolidation of VoIP (Voice over Internet Protocol) technology together with the significant increase of mobile services provided by third- and fourth generation (3G, HSPA, LTE) networks, have recently created the conditions for a considerable expansion of mobile VoIP applications and services. The continuous interest in VoIP applications derives from a number of advantages that this technology offers: cost savings, new value-added services, flexibility and scalability. Mobile IP networks are not designed to support real-time voice traffic because of several drawbacks concerning the wireless medium [1][2][3]: resources sharing, traffic congestion, radio link coverage etc., which impact directly such parameters as delay, jitter, and packet losses. These are the main causes of quality degradation of VoIP calls over the PSTN. While in a fixed network scenario the gap is reduced arbitrarily by an appropriate dimensioning of the characteristics of ADSL access in terms of guaranteed minimum bandwidth or MCR (minimum cell rate), in a cellular network scenario the quality of voice traffic over IP is greatly reduced due to strong current limitations in terms of the requirements regarding delay and guaranteed bandwidth that cannot be arbitrarily decided.

DOI: 10.14738/tnc.32.1157
Publication Date: 25th April, 2015
URL: http://dx.doi.org/10.14738/tnc.32.1157

In recent works the authors have proposed a dual streaming (or packet duplication) approach to mitigate the degradation of speech quality in a scenario of mobile VoIP services over 3G-HSPA. This technique has been previously investigated using simulation [4] and then using real time prototypes [5][6] based on cheap embedded systems. From a series of simulated and experimental measurements based on well-known network metrics (i.e. packet loss rate, end-to-end delay and jitter) and speech quality indexes (i.e. the MOS, Mean Opinion Scores, and PESQ, Perceptual Evaluation of Speech Quality values), it was found out that the perceived quality of communication can be drastically enhanced by sending a duplicate copy of a voice packet exploiting two different network interfaces equipped with USIM belonging to different cellular operators.

As to the costs/benefits balance, the proposed method on the one hand requires a dual RF module, but on the other hand it is also true that nowadays dual-SIM mobile terminals have become very common and allow the implementation of load balancing and fail over mechanisms improving the Internet access for numerous applications. Furthermore, there are contexts where the use of advanced mobile terminals is widely justified to obtain greater continuity and intelligibility of the conversation, i.e. the case of environmental wiretapping for forensic services.

The goal of the present paper is to:

- present the main idea the dual streaming approach is based on;
- describe the hardware/software architecture of a real time prototype implementing it;
- present a simulation study and a complete performance evaluation of the system by using objective metrics;
- evaluate the real impact of the proposed solution on the intelligibility of the transmitted audio signal;
- draw conclusions about the behavior of the system in order to provide some ideas on how to improve weak aspects such as energy power consumption and bandwidth waste by useless packet duplication;
- present the main ideas to enhance further development of the prototype.

The paper is structured as follows: in Section 2 the dual streaming architecture and the hardware/software components of the prototype are described; Section 3 presents the metrics employed to compare traditional single stream transmissions and the Dual VoIP mechanism; Section 4 reports a simulation study performed by using Opnet Modeler as network simulator; in Section 5 the execution of a real test bed, the obtained performance results and the main ideas for the future works are presented; finally, in Section 6 conclusions are drawn.

2 System overview

This paragraph provides a brief overview of both the architecture and the prototype realized with the purpose of evaluating the effectiveness of the suggested approach.

2.1 Motivation and related work

In a 3G/4G network scenario the quality of voice traffic over IP is greatly reduced due to the reliable radio link, delay and guaranteed bandwidth. These are the main causes of quality degradation of VoIP calls. The dual streaming VoIP architecture has been conceived with the aim of enhancing the speech quality and intelligibility of the audio signal. These aspects are very important in common best effort applications using VoIP as well as in particular conditions such as environmental

wiretapping for forensic uses and/or private tactical communications in network-centric contexts where real time listening and intelligibility of the speech signal play a key role.

At the best of our knowledge it is the first work where duplication is performed directly by the source node whereas several solutions have been proposed [7][8][9] involving access routers or network devices belonging to the end-to-end path between the source and destination node. Our solution does not implicate any modification of the network apparatus involving only the interested devices. Last but not least this work presents a deep evaluation of the duplication system that focuses on the intelligibility of the transmitted audio signal not only analyzing well-known networking metrics such as packet loss rate, delay or jitter.

2.2 Architecture

The basic idea of the dual stream approach is very simple and has been described in [4]. With the scope of improving the experience of VoIP users over 3G and HSPA access networks a dual streaming approach was introduced, its primary goal being the reduction of the loss and the delay of VoIP packets by duplicating the single data flow and sending it through two different access networks managed by two different telephone operators. The data flow generated at the application layer was split into two different IP flows, each of which was associated to the related radio interface, in such a way that the source node transmitted two IP flows having different source addresses but the same destination address, i.e. the receiver of the VoIP call. At the receiver side the two data streams were merged in order to deliver a unified data flow to the application layer of the destination. Synchronization was performed by using the sequence number field of the RTP protocol and discarding the duplicate packets received from the two IP sources.

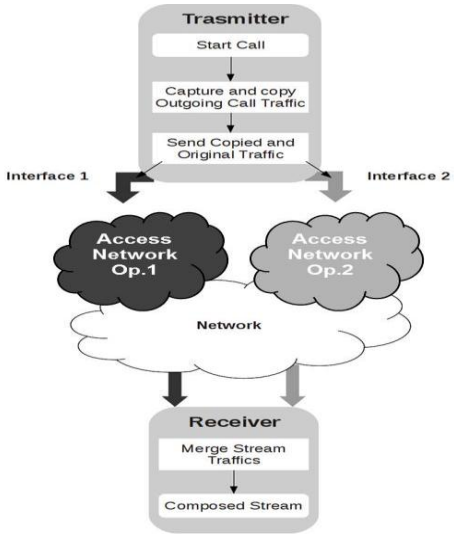


Figure 1. The functional architecture of Dual-VoIP system

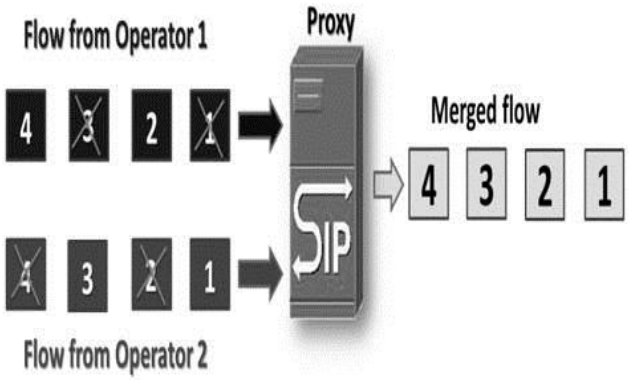


Figure 2. The proxy server at the receiver-side receives the two data flows and merge them in order to provide the upper application layer with a unified merged stream.

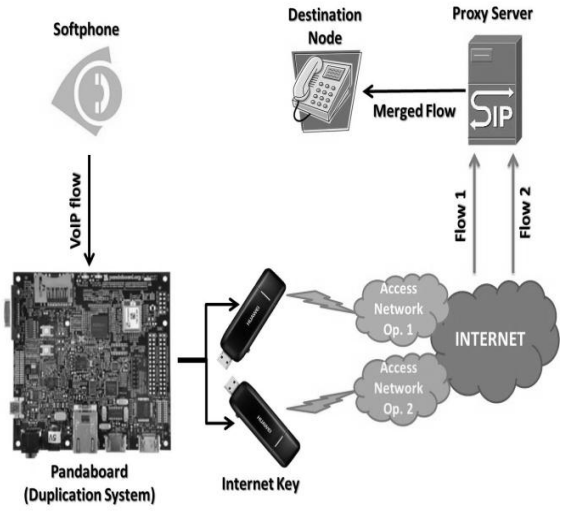


Figure 3. Overview of the Duplication System.

Thus, the workflow of the system can be easily summarized in four steps:

1. Call set up: during this phase SIP negotiation occurs at the end of which IP destination address and RTP destination port are established; furthermore, the transmitting softphone sends the parameters of the flow to the duplication system;
2. Packets duplication: the duplication hardware receives RTP packets from the softphone, duplicates them and finally sends the two obtained flows to the available HSPA network interfaces;
3. Flows merging: at the destination, a proxy server receives the two data flows from the mobile node and stores packets in a buffer discarding duplicates by using the RTP sequence number to identify a copy of the same packet;
4. Reception: the application layer at the destination side receives a single merged VoIP flow from the proxy server.

The functional architecture of the dual streaming transmission is illustrated in Figure 1.

2.3 Description of the prototype

Trivial versions of Dual VoIP prototype have been described in [5] and [6] and consist of a transmitting part and a receiving part. Here we present the final version of Dual VoIP system. The transmitter includes three elements further referred to as T1, T2 and T3:

T1) A softphone realized in Python [10] using PjSIP, PjLib, PjMedia and PjSUA libraries by virtue of which it is possible to establish and manage multimedia communications by using the SIP protocol, negotiating the parameters of the connection between the source and the destination of the VoIP flow; softphone is conceived to communicate with the lower layer, i.e. the duplication system, in order to establish the parameters of the RTP flow to duplicate;

T2) A duplication layer directly implemented in a PandaBoard [11], a low-cost, low-power and open source single-board computer, equipped with a dual core processor ARM Cortex, based on the Texas Instruments OMAP4430 system on a chip (SoC) and running the S.O. Linux pandabord-desktop 3.2 [12];

T3) Two radio interfaces, i.e. two USB internet keys, connected to the PandaBoard, equipped with SIMs belonging to two different network operators.

Softphone and duplication layer communicate according to a client-server paradigm: once established the SIP connection, the softphone sends to the duplication system the IP destination address and the RTP destination port; from here the duplication layer captures the VoIP flow generated by the softphone and duplicates it by sending two copies of the same data flow: the first through the interface 1 and the latter through the interface 2.

To enable the dual streaming technology also the receiver must be modified accordingly. In fact, the receiver has to mix the two data flows received obtaining a single and better data stream that finally is sent towards the listening position. Receiver node has been implemented in a common general porpoise computer and consists of three software elements indicated as R1, R2 and R3:

R1) A proxy server that receives the data flows from the two network interfaces of the transmitter node (data flows have the same destination IP address and the same RTP destination port but have two different IP source addresses related to the two different network interfaces the mobile node is equipped with) and executes the merge of the two streams according to the scheme shown in Figure 2. The merged data flow, obtained by discarding duplicate packets (this feature is performed by exploiting the RTP sequence number of the received packets) is finally sent to the asterisk listening position, a software enabling the real time listening of the VoIP flow;

R2) An Asterisk [13] server that receives the merged flow and sends it to a destination user for the real time listening;

R3) The softphone at the listening position.

An overview of the overall system employed for the test bed under analysis is presented in Figure 3.

3 Evaluation Framework

With the aim of evaluating the proposed mechanism a set of metrics has been selected to compare the traditional transmissions and the Dual VoIP. The metrics taken into consideration for the comparison can be summarized as follows:

Packet loss rate: it indicates the percentage of packets lost due to the bad link quality, network traffic congestion and radio coverage availability; however, it is essential to point out that a packet is considered lost also if it is latecomer, i.e. if the delay between its arrival time and the arrival time of the last accepted packet is greater than the playout buffer implemented at the merger layer of the receiver node;

MOS (Mean Opinion Score): measuring the speech quality is a very hard task because it is influenced by several parameters, i.e. end-to-end latency, jitter, packet loss rate, noise, etc.; for this reason, the MOS index has been introduced with the aim of presenting a universal measure of the speech signal quality. ITU-T P.800 [14] specifies the range of values between 1 and 5, where 1 indicates a very bad signal with continuous interruptions and 5 represents an excellent signal, equal to the AM radio quality. The MOS evaluation procedure of a speech signal includes the calculation of the average score assigned to the latter by a representative sample of population. It is evident that such a method is not objective and expensive, in terms of both economic and temporal efforts;

PESQ (Perceptual Evaluation of Speech Quality): even though the MOS (Mean Opinion Score) index has been introduced with the aim of presenting a universal measure of the speech signal quality, the related evaluation procedure appears to be very complex, thus, the PESQ index has been introduced and will be frequently employed to compare the two approaches. It has been standardized in [15] and represents a full-reference algorithm for evaluating the speech signal quality. It is based on the

comparison between the original speech signal and the one received at the destination node; it is related to the MOS value by the following mapping relationship [16]:

$$MOS = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.4945 \cdot PESQ + 4.6607}}$$

Characterization of cuts: to measure the intelligibility of the received audio trace, particularly useful in case of forensic applications such as environmental wiretapping, it is very important to understand how the packet loss rate and delay can affect the capability of understanding the conversation; with this aim the cuts introduced by wireless link degradation have been classified in four categories based on the length of the cuts to obtain a direct measure of the impact of our solution on this aspect:

a) a) Phoneme: the basic unit of sound in a language which, combined with other phonemes, results in meaningful units. Thus, for the purpose of the present research, it has been assumed that the cuts of phonemes account for approximately 400ms of the entire audio signal;
b) Word: typically one word long cuts within a conversation which account for a period between 400ms and 2 seconds;
c) Sentence: typically one sentence long cuts within a conversation which account for a period from 2 to 30 seconds of the entire audio signal;
d) Talk: cuts more than 30 seconds long within a conversation.

Composition of the merged flow: this index provides a measure of the composition of the merged flow, i.e. the flow received by the softphone at the receiver side, with the scope of establishing how the two original data flows contribute to the production of the dual stream one; if the composition of the merged stream is fair, i.e. each flow contributes to the merged one with a percentage of packets close to 50%, it means that the two operators are contributing equally to the transmission, elsewhere if the composition is unfair - one of the two operators is performing better than the other one. Actually, the prototype is able to measure the metric instantly, and this feature will play a key role for the implementation of a smart stream duplication algorithm to turn on and turn off network interfaces to reduce power consumption when duplication is not needed.

4 Simulation Study

The simulation framework used to evaluate the proposed solution is Opnet Modeler 14.5 [17] jointly with the Wireless Suite [18] for modeling UMTS cellular networks. For the purpose of this paper, we have considered a hypothetical scenario of 6 Km2, where base stations belonging to two different cellular operators, i.e. Operator 1 and Operator 2, were deployed. We suppose that each operator has the same number of base stations equal to NBsOp1 = NBsOp2 = 6, distributed according to the network topology model illustrated in Figure 4, where each base station has a coverage radius equal to 1 Km. As a result, the area of interest is divided into six sub-regions, each of which is situated under the radio coverage area of two base stations, one for each operator. The user equipment, UE, moves within the scenario according to the random waypoint model and establishes, during its movement, a VoIP call with the correspondent node represented by the workstation connected to the IP core network. Furthermore, we suppose that each base station of every single operator is characterized by different signal power strengths and, consequently, different bit error rates (BERs). In such a way, for each sub-region it will be possible to identify a base station providing lower bit error rate and, thus, establish higher quality communication.

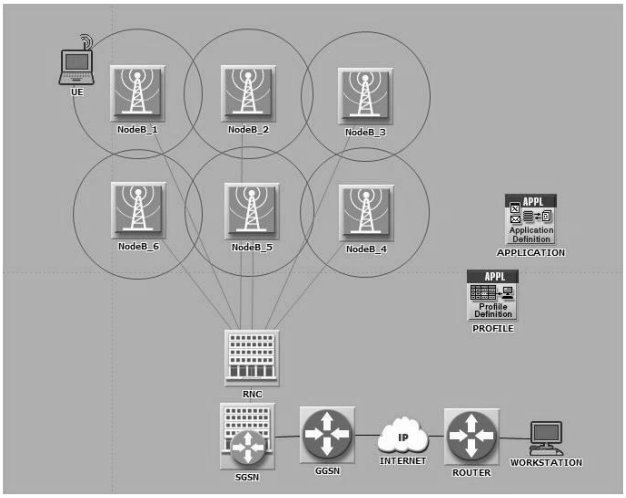


Figure 4: Simulation scenario in Opnet Modeler.

More in detail, we have created three scenarios, each of which is characterized by different packet loss rates. The three cases taken into consideration are:

- Urban: characterized by low-medium BERs, it approximates a scenario where the user equipment operates within a typical urban area;
- Rural: radio coverage is bad and several cells provide a high BER, the UE operates in the countryside;
- Hybrid: characterized by medium BERs, it is used to approximate a situation where the UE operates within a mixed urban-rural environment.

In Figure 5 we illustrate the distribution of the BSs belonging to the two operators, whereas the characteristics in terms of bit error rate for the investigated scenarios are reported in Table I. During our simulations, the UE moves within the area of interest according to the random way point model, with speed uniformly distributed in the range 1-10 m/s and wait time equal to 60 seconds. During the movement, the UE establishes a VoIP call towards the correspondent node, CN, directly connected to the Internet core network. The G.729 codec has been selected as VoIP codec.

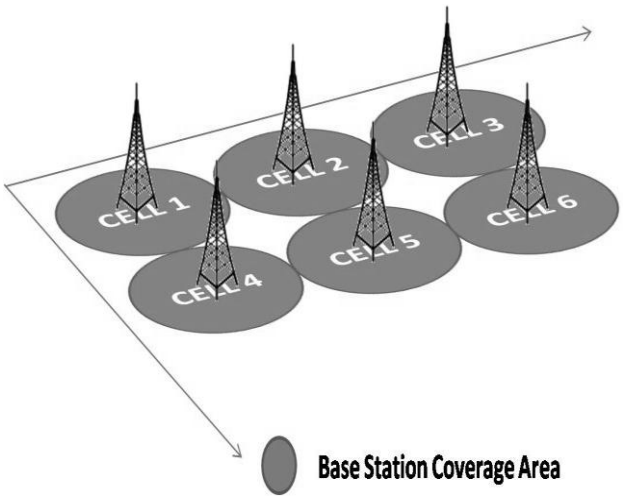


Figure 5: Base Stations distribution for the six-cell scenario

Table 1. BER values guaranteed by the different operators in the six-region scenarios

Cell	Operator 1	Operator 2
BER DISTRIBUTION RELATED TO **URBAN** SCENARIO		
Cell 1	Low BER	Low BER
Cell 2	Low BER	Medium BER
Cell 3	Medium BER	Low BER
Cell 4	Medium BER	Low BER
Cell 5	Low BER	Low BER
Cell 6	Low BER	Medium BER
BER DISTRIBUTION RELATED TO **RURAL** SCENARIO		
Cell 1	Low BER	Low BER
Cell 2	High BER	Medium BER
Cell 3	Medium BER	High BER
Cell 4	Medium BER	High BER
Cell 5	High BER	High BER
Cell 6	High BER	Medium BER
BER DISTRIBUTION RELATED TO **HYBRID** SCENARIO		
Cell 1	Low BER	Low BER
Cell 2	Low BER	Medium BER
Cell 3	Medium BER	High BER
Cell 4	Medium BER	Low BER
Cell 5	High BER	High BER
Cell 6	High BER	Medium BER

4.1 Simulation Results

Firstly, we investigated the impact of the proposed approach on the packet loss rate. With this aim, we performed 20 runs for each scenario measuring the packet loss rate for the two different operators and the dual streaming approach. Results are briefly summarized in Table 2, where it can be observed that the use of the proposed mechanism drastically reduces the number of packet lost in the three investigated scenarios. More in detail, in a typical urban scenario, characterized by lower BER compared with the other scenarios, we can observe that packet loss rate is reduced up to 89% respect with the single operators. In rural scenario, subjected to a very high BER, Dual VoIP outperforms single operators by reducing the packet loss rate of nearly 60%. The hybrid scenario shows the behaviors of single operators and Dual VoIP in case of intermediate BER in a mixed environment. Also in this case we can note a significant performance improvement due to the use of the replicated stream.

In order to evaluate the proposed approach in terms of PESQ and MOS indexes, we used an audio book, divided into ten sub-sections, each lasting four minutes, to simulate the payload of stream suffering packet loss in the proposed scenarios. We applied cuts to the original audio stream coherently with the packets lost during the simulations by the two operators and the dual streaming technique. Finally, rebuilt traces together with the original one have been analyzed obtaining PESQ and MOS scores for the two operators' streams, as well as the dual stream.

Results presented in Figure 6, show the effectiveness of the proposed technique in terms of perceived speech quality for the three different scenarios: PESQ and MOS scores are greatly improved by using the dual streaming technique, guaranteeing more than one PESQ score point in the investigated scenarios.

As a conclusion, we can affirm that the quality of conversation can be considerably improved by using the dual streaming approach. In fact, MOS and PESQ indexes of single operators range from very low to medium-high scores, which means that, in several cases, the intelligibility and comprehension of human voice in a conversation is unsatisfactory or even completely incomprehensible. Dual-streaming supplied calls grant an enhancement in terms of speech quality, reducing situations in which talks became not understandable.

Table 2. Packet loss decrease ratio.

	URBAN	RURAL	HYBRID
$PLD_{Operator\ 1}$	86,80%	59,69%	75,84%
$PLD_{Operator\ 2}$	89,58%	57,75%	74,64%

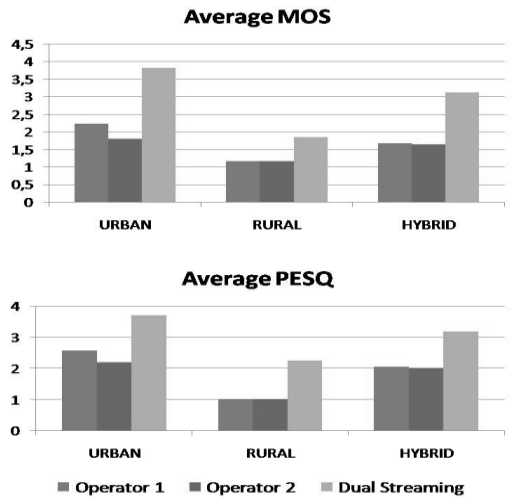


Figure 6: Average values of MOS and PESQ indexes for the different scenarios

5 Performance Evaluation of the Prototype

The simulation study presented in the above section permits us to obtain an approximate evaluation of the effectiveness of the proposed approach because the scenarios, i.e. the position, the coverage area and performances in terms of bit error rate of the cellular base stations, are hypothetical. With the aim of obtaining a more detailed and realistic evaluation of the Dual VoIP architecture a real test bed is presented in this section.

5.1 Description of the test bed

Tests consist in 50 VoIP phone calls made with the previously described architecture. Each call lasted from 2 to 4 minutes; calls have been made under different conditions, in terms of movement speed and mobile phone coverage. To explore the response of the system in different network conditions, tests were carried out moving the prototype in urban and countryside areas, in indoor and outdoor scenarios, and investigating both pedestrian and vehicular movement conditions. This permitted us to check the effectiveness of the dual streaming system under good and adverse radio coverage conditions. The mobile transmission system started a VoIP call towards a fixed destination node, i.e. the listening position, which is the proxy server represented in Errore. L'origine riferimento non è stata trovata. at the receiver side. At this point we employed Wireshark [19] to calculate the packet loss rate at each network interface (single streams) of the proxy server and at the destination node (merged flow), and ChanSpy [20] to wiretap the VoIP flows at each interface of the proxy server and at the destination node in order to evaluate the PESQ index for single stream and dual stream

techniques. Two tools at the destination side permitted us to calculate the number cuts, classify them according their length (i.e. the number of sequential packets lost), and establish the composition of merged flow.

5.2 Results

In this section the results obtained by the prototype are illustrated and the conclusions about the behavior of the duplication system are drawn; moreover, possible improvements to the existing system are provided, which will be the aim of our future works.

First of all, two versions of receiver were considered for the purpose of the present research, based on the presence or not of a play out buffer in the merging block. We introduce this element in order to make the system independent of the upper application layer, that could have or not a proper play out buffer, and put in order the packets coming from the two source IP addresses that, obviously, can be affected by different end-to-end delay and jitter according to the access network conditions. Results are presented in Figure 7 and Figure 8 where we can note how the presence of a play out buffer at the merger layer enhances the perceived speech quality at the destination node. According to this consideration, from now on we will consider the performance evaluation of the receiver with play out buffer. In particular, for our test bed a playout buffer of 200ms has been selected.

Figure 9 and Figure 10 report the packet loss rate and PESQ values related to the two different investigated scenarios, i.e. urban and countryside. Dual VoIP prototype improves significantly both parameters, obtaining a reduction of packet loss rate up to 90% and a PESQ enhancement up to 1 points. The performance improvement is more remarkable in the case of countryside scenario because of the frequent lack of radio coverage provided by 3G/HSPA operators. This conclusion permits us to affirm that Dual VoIP represents a viable solution to improve real time transmissions in the context of vehicular communications.

Regarding the audio clipping (that represents a key parameter in network-centric and forensic services such as tactical communications and environmental wiretapping respectively) due to a packet loss during the calls, the Dual VoIP outperforms the single stream transmissions in terms of absolute number of cuts and, more importantly, the implementation of Dual VoIP significantly impacts the statistical distribution of the cuts by drastically reducing the longer cuts, i.e. words, sentences and talks. Figure 11 and Figure 12 show the enhancements introduced by using our approach as opposed to the common single stream transmissions.

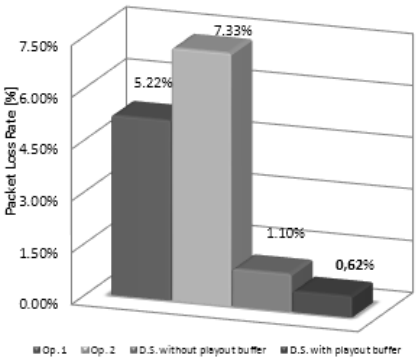


Figure 7. PLR comparison of the two operators and the dual streaming prototype with and without the play out buffer

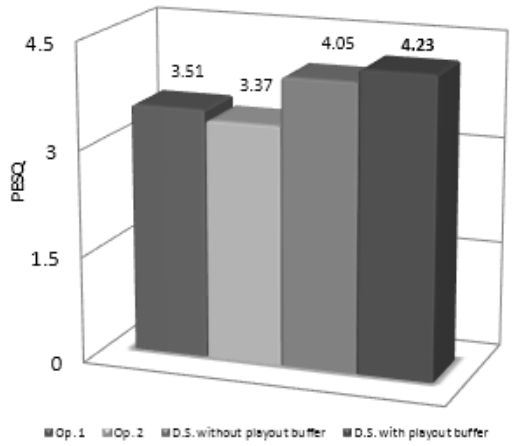


Figure 8: PESQ evaluation with and without the play out buffer

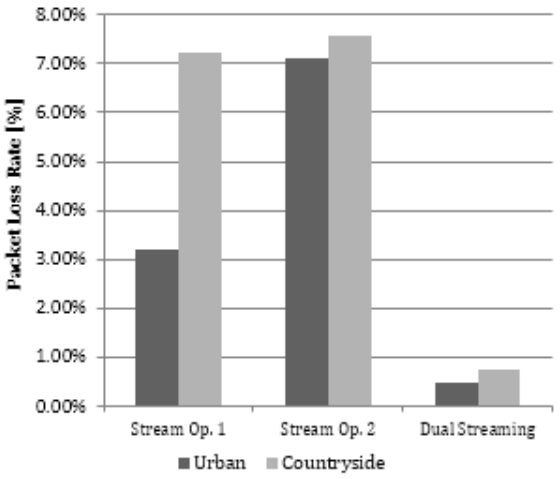


Figure 9: PLR measured in urban and countryside scenarios

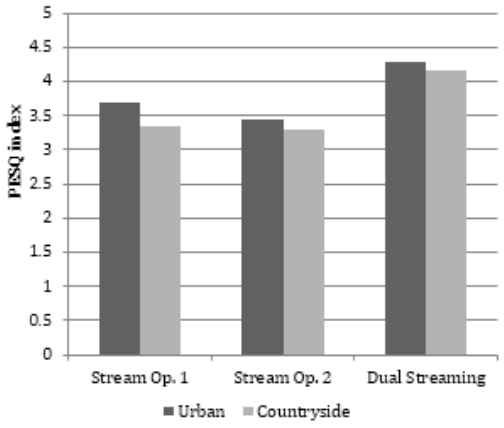


Figure 10. PESQ index measured in urban and countryside scenarios

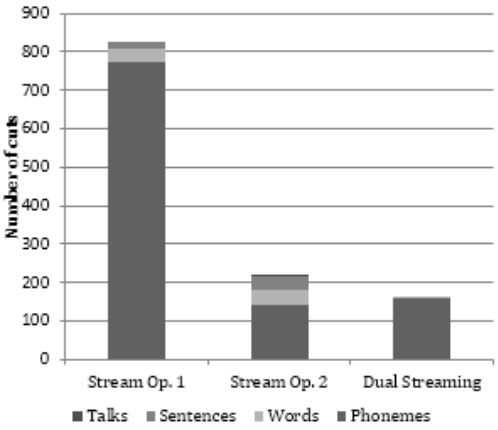


Figure 11. Number of cuts estimated during the tests, classified according to their length

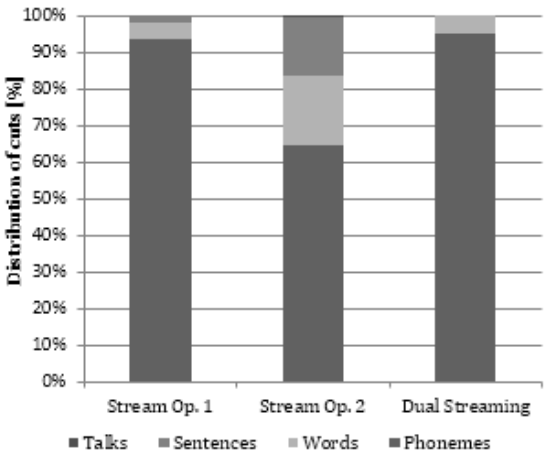


Figure 12. Statistical distribution of the cuts

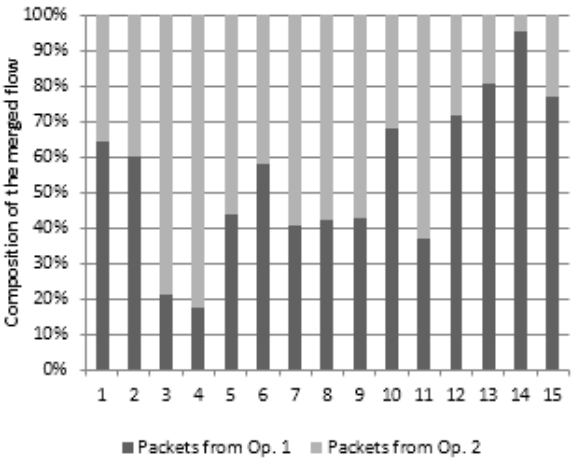


Figure 13. Statistics about the composition of the merged dual stream flows for 15 VoIP calls in urban scenarios

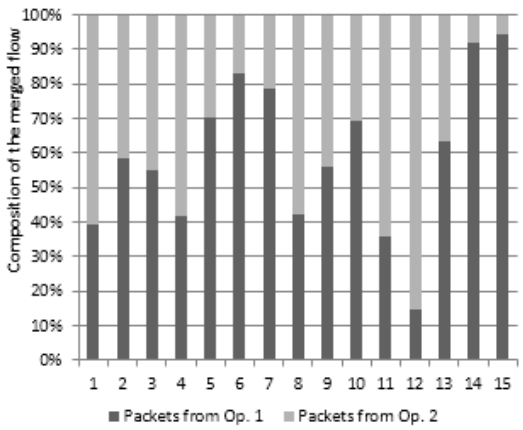


Figure 14. Statistics about the composition of the merged dual stream flows for 15 VoIP calls in countryside scenarios.

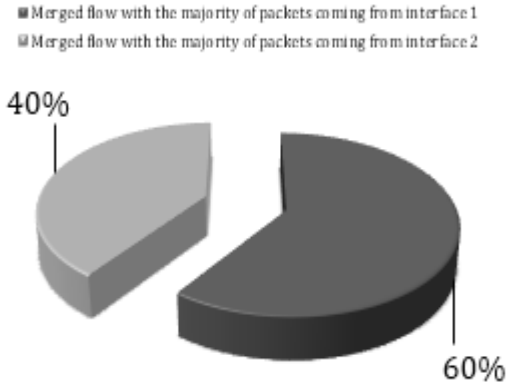


Figure 15. Statistics about the composition of the merged dual stream flows at the listening position

Finally, Figure 15 give a measure of the composition of the merged flow at the output of the play out buffer of the merger layer. This parameter allows us to establish how the two network interfaces at the transmitting side contribute to the composition of the dual stream flow at the listening position. The percentage reported in the figure indicates that one interface has performed better than the other during the test but, at the same time, the difference of contribution is not large enough to justify the use of only one interface instead of the other. During the realization of the prototype we also included a real time measure of this parameter in order to establish, in the future version of the prototype, if it is possible to turn off one of the two interfaces at any time, and thus work in a single stream mode, or it is compulsory to work in dual mode by using both interfaces at all times.

5.3 Case studies

In this section two particular cases observed during the test performed by using the prototype are reported. In the first case, the dual streaming operates as a backup solution: the radio link of one of the two cellular operators goes down but the speech signal at the listening position is received correctly; as we can see in the wave trace showed in Figure 16, the packets lost by one operator are retrieved from the other connection and vice versa. When such a case appears, as a future work, in the next version of the prototype we will introduce a smart interface selection algorithm with the aim of reducing power consumption by turning off the transmitting interface that does not contribute to the merged flow at the receiver. In this regard, Table 3 reports the metrics registered when such a case arises. The second case, let us define it the failover case, takes place when both operators lose packets almost equally due to bad links quality; if this is the case the merged flow is constituted by packets coming from the two IP source addresses and packets lost in the first connection are recovered by the other and vice versa. Obviously, when fail over mode is detected, the above mentioned smart selection of the transmitting interface cannot be enabled. It has to be admitted, however, that the algorithm for the adaptive selection of the interface and/or the switching between the dual and single stream transmission is not the object of the present paper. Figure 17 and TABLE 4 refer to the above mentioned case.

Table 3. Backup Mode

Metric	Stream		
	Stream Op. 1	*Stream Op. 2*	*Dual Streaming*
P.L.R.	0.00%	46.52%	0.00%
PESQ	4.5	0.39	4.5
Characterization of cuts			
Phonemes	0	5	0
Words	0	1	0
Sentences	0	3	0
Talks	0	1	0
Composition of the flow			
Pkts from Op.1	100	-	94.43%
Pkts from Op.2	-	100	5.57%

Table 4. Failover Mode

Metric	Stream		
	Stream Op. 1	*Stream Op. 2*	*Dual Streaming*
P.L.R.	19.66%	21.58%	0.17
PESQ	1.88	0.94	3.89
Characterization of cuts			
Phonemes	34	7	2
Words	3	1	0
Sentences	1	3	0
Talks	0	0	0
Composition of the flow			
Pkts from Op.1	100	-	69.28%
Pkts from Op.2	-	100	30.72%

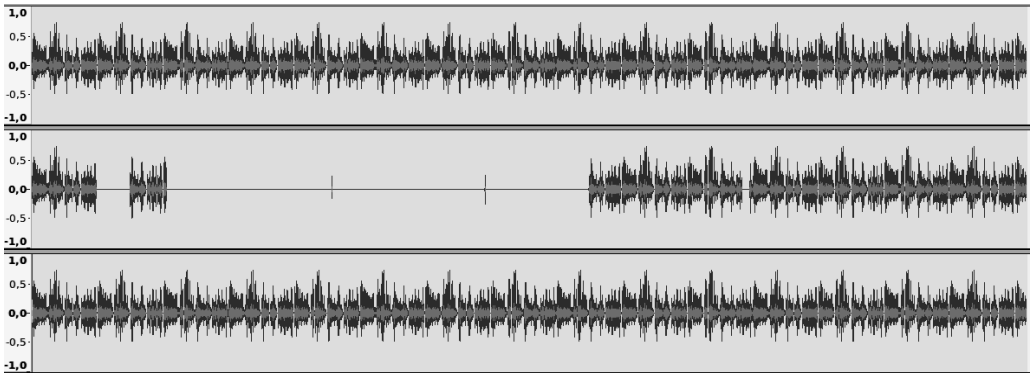


Figure 16. Waveforms observed during the backup mode. From the top to the bottom the audio trace of Operator 1, Operator 2 and Dual VoIP

Figure 17. Waveforms observed during the failover mode. From the top to the bottom the audio trace of Operator 1, Operator 2 and Dual VoIP

5.4 Prototype enhancement

The results presented in the above section permit us to affirm that the Dual VoIP system drastically impacts on the speech quality and intelligibility of the received signal. Nevertheless, we are aware that the proposed architecture presents two main drawbacks; first of all the prototype needs to supply two network interfaces during its working and as second aspect it implies a waste of bandwidth due to useless packet duplication when one interface should be sufficient to guarantee an efficient audio signal transmission. With the aim of counteract the above mentioned aspects we are working on the deployment of a new prototype, let us define it as Opportunistic Dual VoIP mechanism, enabling a smart use of the radio access facilities. More in detail, the new version will include a new functional block at the control layer, the flow manager, at both transmitter and receiver side. The two flow manager communicate in order to establish if the mobile device has to operate in single stream or in dual stream mode. When single stream mode operates the stream manager should give information to the duplication layer about the network interfaces to be used.

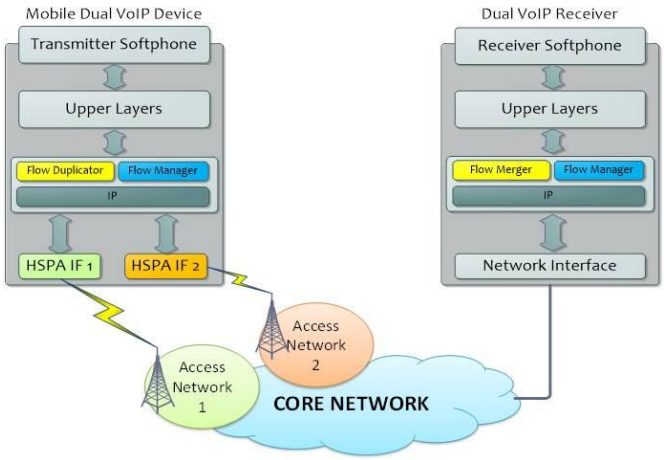


Figure 18. The protocol stack of the enhanced prototype

The algorithm at the base of flow manager is out of the scope of the present paper but for the aim of clarity we can say that it will operate according to information related to:

a) Operation mode of network interfaces (3G, HSPA, LTE, etc.);
b) Coverage area of the near base station (signal level perceived at the transmitter);
c) Bandwidth constrains (the USIMs the device is equipped with could have limitations in terms of data, time, etc.);

d) Real time performance evaluation at the receiver side: the flow manager at the receiver knows how the two source interfaces are contributing to the merged flow having measures about delay, jitter, packet loss rate and composition of the merged flow.

To define the parameters to take into consideration and the observing window duration at the receiver side will comport an effort in terms of analytical and simulation studies before to release a new version of the smart prototype but we are trusting in the capability of drastically reducing the power consumption of the device at the cost of a slight reduction in speech quality and intelligibility. The new version, furthermore, will be able to manage other multimedia contents such as real time video. Future applications will concern:

- Real time audio and video communications from mobile and or vehicular devices;
- Communications in network-centric environments;
- Forensic services and applications such as environmental wiretapping;
- Vehicle-to-vehicle and vehicle-to-network data transmissions.

6 Conclusion

The present paper provides an extended performance evaluation of a mobile VoIP duplication system conceived to enhance several applications such as VoIP transmission over 3G-HSPA networks, HD VoIP services in mobile scenarios, VoIP environmental wiretapping for forensic applications, tactical communications in network-centric applications and others. Results obtained by the means of simulations and, more importantly, by the means of a real-time prototype indicate the effectiveness of the proposed approach in the reduction of the packet loss rate, the number of cuts the speech signal is affected by, and, last but not least, improving the PESQ index. In particular, the Dual VoIP technique significantly influences the statistical distribution of the cuts by drastically reducing longer cuts, i.e. words, sentences and talks. On the other hand, we are aware that the prototype has the drawback of power consumption due to the need to supply two network interfaces simultaneously. We are currently working on a solution to solve this problem by introducing an enhancement of the prototype by implementing an adaptive and smart use of the interfaces based on real-time evaluation of different connections. In particular, we will consider a mobile observation window during which the receiver measures packet loss rate, jitter and composition of the merged flow in order to enable the use of the common single stream transmission or the dual streaming mode at the transmitter side. This feature will be particularly useful in case of forensic applications and services where the power autonomy of the device play a key role. Finally, additional enhancement of the prototype will concern the use of heterogeneous network interfaces such as 3G-HSPA, LTE and WiMAX.

REFERENCES

[1]. A. Lamba, J. Yadav, G. U. Devi, “Analysis of Technologies in 3G and 3.5G Mobile Networks”, in Proc. of Int. Conf. on Communication Systems and Network Technologies (CSNT), 11-13 May, 2012, pp. 330-333.

[2]. S. Jadhav, Haibo Zhang, Zhiyi Huang, “Performance Evaluation of Quality of VoIP in WiMAX and UMTS”, in Proc. of IEEE 12th Int. Conf. on Parallel and Distributed Computing, Applications and Technologies (PDCAT), 2011.

[3]. A. Takahashi, H. Yoshino, and N. Kitawaki, "Perceptual QoS assessment technologies for VoIP", IEEE Communications Magazine, vol. 42, no. 7, pp. 28-34, July 2004.

[4]. F. Beritelli, A. Gallotta, C. Rametta, "A Dual Streaming Approach for Speech Quality Enhancement of VoIP service over 3G networks", Proceedings of IEEE Digital Signal Processing (DSP), 1-3 July 2013.

[5]. F. Beritelli, A. Gallotta, S. Palazzo, C. Rametta, "Dual stream transmission to improve mobile VoIP services over HSPA: a practical test bed", Proceedings of IEEE 8th International Symposium on Image and Signal Processing and Analysis (ISPA), 4-6 September 2013.

[6]. F. Beritelli, C. Rametta, "HSDPA Dual Streaming Approach for Improving VoIP Speech Quality in Forensic Applications", Proceedings of IEEE 9th International Symposium on on Communications Systems Networks and Digital Signal Processing (CSNDSP) 2014, 23-25 July 2014, Manchester, UK.

[7]. Dennis C. Ferguson, "Data Duplication for transmission over computer networks", U.S. Patent 7342890, March 11, 2008.

[8]. Niall Thomas Davidson, "Low delay lossless packet selector", U.S. Patent 2012/0106330, May 3, 2012.

[9]. P. R. Michaelis, R. Toennis, D. M. Grover, "System and method for providing a replacement packet", U.S. Patent 2010/0188967, July 29, 2010.

[10]. http://www.python.org/.

[11]. http://pandaboard.org.

[12]. http://lubuntu.net/.

[13]. http://www.asterisk.org/.

[14]. http://www.itu.int/rec/T-REC-P.800/en.

[15]. http://www.itu.int/rec/T-REC-P.862/.

[16]. http://www.itu.int/rec/T-REC-P.862.1/en.

[17]. http://www.opnet.com/solutions/network_rd/modeler.html.

[18]. http://www.opnet.com/solutions/network_rd/modeler_wireless.html.

[19]. http://www.wireshark.org.

[20]. http://www.asteriskdocs.org/en/2nd_Edition/asterisk-book-html-chunk/asterisk-APP-B-351.html

SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
TNC Transactions on Networks and Communications
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

Classification of Web Services using Fuzzy Classifiers with Feature Selection and Weighted Average Accuracy

V. Mohan Patro[1] and Manas Ranjan Patra[2]
Department of Computer Science, Berhampur University, Berhampur, Odisha, India
[1]vmpatro@gmail.com, [2]mrpatra12@gmail.com

ABSTRACT

Web services have become an innovative and accepted means of service delivery over the Internet. In recent years there has been astounding growth in the number of web services provisioned by businesses and corporate houses. In the presence of a plethora of web services, a service consumer faces the real challenge of making a right choice based on certain preferences. Therefore, it becomes necessary to classify a set of web services based on certain quality parameters in order to facilitate user choice of web services under different scenarios. Several classification techniques have been proposed by researchers to classify data sets in different application domains. In this work, we have employed three fuzzy classifiers, namely, Fuzzy Nearest Neighbor, Fuzzy Rough Nearest Neighbor, and Fuzzy Rough Ownership Nearest Neighbor to classify web services. We have used the standard QWS dataset for our experimentation. The accuracy of the classifiers has been computed with and without feature selection. In order to further improve classification accuracy, a Weighted Average Accuracy technique has been applied to the confusion matrix obtained after feature selection.

Keywords – Web services, Fuzzy Nearest Neighbor classifier, Fuzzy Rough Nearest Neighbor classifier, Fuzzy Rough Ownership Nearest Neighbor classifier, Weighted Average Accuracy.

1 Introduction

Web Services are emerging technologies that enable machine-to-machine communication and reuse of services over the Web. A Web Service is a software function provided at a network address and can support interoperable machine-to-machine interaction over the web. Different software systems often need to exchange data with each other, and a web service is a means of communication that allows two software systems to exchange data over the internet. With the increasing number of available Web services on the internet, Web service discovery becomes a challenging issue. It is time consuming to traverse the whole of the Internet with a view to find a Web service that matches one's service requirements. To speed up service discovery, classification can be a useful approach. Researchers have applied different classification techniques to categorize web services based on a set of quality parameters.

Yuan-jie et al. in [1] applied automatic web service semantic annotation and use three classification methods, namely, Naïve Bayes, SVM and REPTree along with ensemble learning. They applied 10 cross-validations of Naïve Bayes, SVM, REPTree and AdaBoost on WSDL files. According to the experiment done on 951 WSDL files and 19 categories, the highest accuracy was 87.39%.

Web Service is an innovative mechanism for rendering services over diversified environment [2]. Efficient result has been taken from QWS dataset using weka tool in the experiment. The experiment results shown in the study are about classification accuracy obtained by J48 as 63%.

DOI: 10.14738/tnc.32.1140
Publication Date: 28th April, 2015
URL: http://dx.doi.org/10.14738/tnc.32.1140

Authors in [3] developed various classification models based on intelligent techniques namely BPNN, PNN, GMDH, TreeNet, CART, SVM and J48 to predict the quality of a web service based on a number of QoS attributes. They observed that J48 out performs other classifiers they used for accuracy calculation.

In [4], authors have shown how SVM is helpful in the classification of web services. They used the SVM (Support Vector Machine) text classification algorithm to classify the service documents based on a standard and widely used taxonomy with feature selection.

Mohanty et al. in [5] employed Naïve Bayes, Markov blanket and Tabu search techniques to classify web services dataset. They noted that the average accuracy of Naïve Bayes classifier is 85.62%, followed by Tabu search of 82.45% and Markov blanket of 81.36%.

In this paper, we have employed three classifiers, namely, Fuzzy Nearest Neighbor (FNN), Fuzzy Rough Nearest Neighbor (FRNN), and Fuzzy Rough Ownership Nearest Neighbor (FRONN) to classify web services dataset. The classification accuracies of the classifiers have been evaluated with and without feature selection. Next, a Weighted Average Accuracy algorithm (from our earlier work [6]) is applied to the confusion matrix obtained after feature selection in order to improve upon the results.

The rest of the paper is organized as follows: Section 2 describes the classifiers used, section 3 presents the experimental set up, section 4 analyzes the results, and section 5 concludes the paper.

2 Classification Techniques Used

2.1 Fuzzy Nearest Neighbors

The Fuzzy Nearest Neighbor (FNN) algorithm [7, 8] was introduced to classify test objects based on their similarity to a given number K of neighbors, and these neighbors' membership degree to (crisp or fuzzy) class labels. For the purpose of (FNN), the extent C(y) to which an unclassified object y belongs to a class C is computed as:

$$C(y) = \sum_{x \in N} R(x,y)C(x) \quad (1)$$

where N is the set of object y's K nearest neighbors, and R(x,y) is the [0,1]-valued similarity of x and y.

The Fuzzy K-Nearest Neighbors Algorithm

FNN (X, *C*, y, K)

Input: X, the training data set; *C*, the set of decision classes;
y, the objects to be classified; K, the number of nearest neighbors.

begin

N ← get Nearest Neighbors (y, K)

For each C ∈ *C* do

C (y) = $\sum_{x \in N} R(x,y)C(x)$

output: $\arg\max_{C \in C}$(C (y))

end

2.2 Fuzzy-Rough Nearest Neighbor

In Fuzzy-Rough Nearest Neighbor (FRNN) algorithm the nearest neighbors are used to construct the fuzzy lower and upper approximations of decision classes, and test instances are classified based on their membership to these approximations. FRNN algorithm combines fuzzy-rough approximations with the classical FNN approach [7,8]. The rationale behind the algorithm is that the lower and upper approximation of a decision class, calculated by means of the nearest neighbors of a test object y, provides good clues to predict the membership of the test object to that class. The algorithm is dependent on the choice of a fuzzy tolerance relation R. Given the set of conditional attributes A, the fuzzy tolerance relation R is defined by

$$R(x,y) = \min_{a \in A} R_a(x,y) \tag{2}$$

in which Ra (x,y) is the degree to which objects x and y are similar for attribute a. Here we choose

$$R_a(x,y) = 1 - \frac{|a(x)-a(y)|}{|a_{max}-a_{min}|} \tag{3}$$

If (R ↓ C) (y) is high, it reflects that all of y' s neighbors belong to C. A high value of (R ↑ C) means that at least one neighbor belongs to that class.

The Fuzzy Rough Nearest Neighbors Algorithm:

FRNN (X, C, y)
X, the training data set; C, the set of decision classes;
y, the object to be classified;
begin
N ← get Nearest Neighbors (y, K)
τ ← 0, Class ← Ø
for each C ∈ C do
if ((R↓C)(y) + (R↑C)(y)) / 2 ≥ τ then
τ ← ((R↓C)(y) + (R↑C)(y)) / 2
end
end
output Class
end

2.3 Fuzzy Rough Ownership Nearest Neighbors

Fuzzy-Rough ownership is an attempt to handle both "fuzzy uncertainty" and "rough uncertainty"[7, 8]. The fuzzy-rough ownership function τ_c of class C is defined for an object y as,

$$\tau_c(y) = \sum_{x \in X} \frac{R(x,y)C(x)}{|X|} \tag{4}$$

The fuzzy relation R is determined by

$$R(x, y) = \exp(-\sum_{a \in A} K_a(a(y) - a(x))^{2/(m-1)}) \tag{5}$$

where, m controls the weighting of the similarity and Ka is a parameter that decides the bandwidth of the membership. Ka is defined as

$$K_a = \frac{|X|}{2\sum_{x \in X} ||a(y)-a(x)||^{2/(m-1)}} \tag{6}$$

$\tau_c(y)$ is interpreted as the confidence with which y can be classified to class C. The algorithm does not use fuzzy lower or upper approximations to determine class membership.

The Fuzzy Rough Ownership Nearest Neighbors Algorithm:

FRONN(X, A, C, y)

X, the training data set; A the set of conditional features;
C, the set of decision classes; y the object to be classified

begin

for each a $\in$ A do

$$K_a = \frac{|X|}{2\sum_{x \in X} ||a(y)-a(x)||^{2/(m-1)}}$$

end

$N \leftarrow |x|$

for each $C \in C$ do $\tau_c(y) = 0$

for each x $\in$ N do

$$d = \sum_{a \in A} Ka(a(y) - a(x))^2$$

for each C $\in$ C do

$$\tau_c(y) += C(x).\exp(-d^{1/(m-1)}) / |N|$$

end

end

end

output $\arg\max_{C \in C}(C(y))$

end

3 Experimentation

3.1 Data Set

The QWS (Quality of Web Service) dataset [9-11] consists of data from over 5000 web services out of which the public dataset consists of a random 364 web services. The service descriptions were collected using the Web Service Crawler Engine (WSCE) [12]. The majority of Web services were obtained from public sources on the Web including Universal Description, Discovery, and Integration (UDDI) registries, search engines, and service portals. The public dataset consists of 364 web services each with a set of nine Quality of Web Service (QWS) attributes that have been measured using commercial benchmark tools. WSRF is used to measure the quality ranking of a web service based on the nine quality parameters (1-9 in Table-1).

In table 1, the service parameters 1-9 are used for computation of classification accuracy with respect to four "Service Classification" values, namely, "Platinum" (high quality), "Gold", "Silver" and "Bronze" (low quality) equivalent to 1 through 4 respectively.

Table 1: QWS Parameter description

P-ID	Parameter Name	Description	Units
1	Response Time	Time taken to send a request and receive a response	ms
2	Availability	Number of successful invocations/total invocations	%
3	Throughput	Total Number of invocations for a given period of time	Invokes per second
4	Success ability	Number of responses / number of request messages	%
5	Reliability	Ratio of the number of error messages to total messages	%
6	Compliance	The extent to which a WSDL document follows WSDL specification	%
7	Best Practices	The extent to which a Web service follows WS-I Basic Profile	%
8	Latency	Time taken for the server to process a given request	ms
9	Documentation	Measure of documentation (i.e. description tags) in WSDL	%
10	WSRF	Web Service Relevancy Function: a rank for Web Service Quality	%
11	Service Classification	Levels representing service offering qualities (1 through 4)	Classifier
12	Service Name	Name of the Web service	None
13	WSDL Address	Location of the Web Service Definition Language (WSDL) file on the Web	None

3.2 WEKA Workbench

We have used the WEKA (Waikato Environment for Knowledge Analysis) machine learning platform [13] for our experimentation. The WEKA workbench consists of a collection of implemented popular learning schemes, which can be used for practical data mining and machine learning.

3.3 Cross-Validation

Cross-validation calculates the accuracy of the model by separating the data into two different subsets, namely, training set and validation set or testing set. The training set is used to perform the analysis and the validation set is used to validate the analysis. This validation process is continued k times to complete the k-fold cross validation procedure. We have used 10-fold cross-validation wherein the dataset is partitioned into 10 subsets, of which 9 subsets are used as the training fold and the 10thsubset is used for testing. The process is repeated 10 times such that each subset is used as a test subset once. The estimated accuracy is the mean of the estimates for each of the classifiers.

3.4 Feature Selection (FS)

Feature selection refers to the process of selecting relevant attributes and reducing redundant and irrelevant attributes in the dataset to improve upon classification accuracy. Therefore, suitable attribute selection method for selecting the most prominent features (attributes) from the dataset is of paramount importance to enhance the performance of classification accuracy and reduce the computation time. In this study, we have applied two feature selection techniques, namely, Information Gain Attribute Evaluator and Gain Ratio Attribute Evaluator.

3.4.1 Information Gain (IG)

It evaluates the worth of an attribute by measuring the information gain with respect to a class. Information gain measure is used to determine how accurately a particular attribute classifies the training data. Information gain is based on the concept of entropy which is widely used in the Information theory domain.

Let node N represents the tuples of partition D. The attribute with the highest information gain is chosen as the splitting attribute for node N. This attribute minimizes the information needed to classify tuples in the resulting partitions and reflects the least randomness or impurity in these partitions [14].

The expected information needed to classify a tuple in D is given by

$$\text{Info (D)} = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (7)$$

where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_{i,D}|/|D|$. Info(D) is the average amount of information needed to identify the class label of a tuple in D.

$$\text{Info}_A\text{ (D)} = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(D_j) \quad (8)$$

The term $\frac{|D_j|}{|D|}$ acts as the weight of the j-th partition. $\text{Info}_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A. Information gain is defined as the difference between the original information requirement and new information requirement. That is

$$\text{Gain(A)} = \text{Info(D)} - \text{Info}_A(D) \quad (9)$$

Using Information Gain Evaluation with Ranker Search on the QWS data set, top 4 attributes (WSRF, WSDL Address, Service Name and Reliability) are selected for classification.

3.4.2 Gain Ratio (GR)

It evaluates the worth of an attribute by measuring the gain ratio with respect to the class.It applies a kind of normalization to information gain using a "split information" value. The split information value represents the potential information generated by splitting the training data set D into v partitions corresponding to v outcomes on attribute A, and is expressed as [14]:

$$\text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \quad (10)$$

The gain ratio is defined as

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)} \quad (11)$$

The attribute with the maximum gain ratio is selected as the splitting attribute.

Using Gain Ratio Evaluation with Ranker Search on the QWS data set, top 5 attributes (WSRF, Throughput, Response Time, Reliability and WSDL Address) are selected for classification.

3.5 Confusion Matrix

Table 2: Confusion Matrix (2×2)

		Predicted Class	
		C_1	C_2
Actual Class	C_1	True positive	False negative
	C_2	False positive	True negative

C_1 – particular class C_2 – different class

True positive (TP) - The number of instances correctly classified as C1
True negative (TN) - The number of instances correctly classified as C2
False positive (FP) - The number of instances incorrectly classified as C1 (actually C2)
False negative (FN) - The number of instances incorrectly classified as C2 (actually C1)
Using the above the following performance parameters are computed:
TP rate (TPR, Sensitivity, Recall) = TP / (TP + FN)
Positive Predictive Value (PPV, Precision) =TP / (TP + FP)
False Discovery Rate (FDR) = FP / (FP + TP)
FP Rate (FPR, False Alarm Rate (FAR), Fall-out) = FP / (FP + TN)
TN Rate (TNR, Specificity (SPC)) = TN / (TN + FP)
Negative Predictive Value (NPV) = TN / (TN + FN)
False Omission Rate (FOR) = FN / (FN + TN)
FN Rate (FNR) = FN / (FN + TP)
Accuracy (ACC) = (TP + TN) / (TP + FN +TN + FP)
F-Value = (2 × Precision × Recall) / (Precision + Recall)

4 Result and Discussion

Here, we analyze the performance of three classification techniques, viz. Fuzzy Nearest Neighbor (FNN), Fuzzy Rough Nearest Neighbor (FRNN) and Fuzzy Rough Ownership Nearest Neighbor (FRONN)along with two feature selection techniques, namely Information Gain (IG) and Gain Ratio (GR). Performance is also observed with the application of weighted Average Accuracy (WAA) [6]. The classifiers are tested using10-fold cross validation.

Table 3: Comparision of classification accuracy(in %) without WAA

Classifier	Without FS	With GR	With IG
FRNN	77.4725	88.4615	93.1319
FNN	80.4945	88.1868	93.6813
FRONN	**86.8132**	**91.2088**	**96.978**

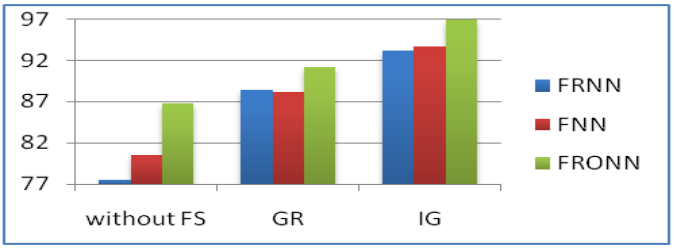


Figure 1: Comparative analysis of classification accuracies(in %) without WAA

Classification accuracy values are recorded in table 3. Classification accuracy values obtained with feature selection give better result than that obtained without feature selection. Accuracy with the use of information gain as feature selection is more than that of gain ratio. From table 3 and fig 1 it is clear that Fuzzy Rough Ownership NN classification technique provides better accuracy as compared to other techniques and the value for Information Gain feature selection is best.

Table 4: Comparision of classification accuracies with WAA

Classifier	without FS	With GR	With IG
FRNN	0.870871573	0.934669726	0.959734633
FNN	0.889362698	0.933628185	0.963991366
FRONN	**0.924103369**	**0.948934307**	**0.982150405**

After applying weighted average accuracy, it is observed that Fuzzy Rough Ownership NN classification technique provides better accuracy as compared to other techniques and the value for Information Gain feature selection is again the best (Table 4 and Fig 2).

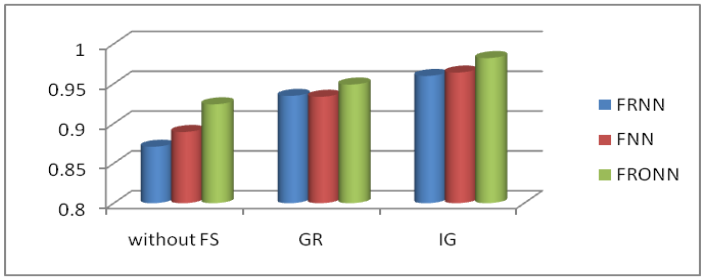


Figure 2: Comparative analysis of classification accuracies with WAA

Comparing the results obtained in table 3 with that of table 4, we found that the weighted average accuracy values are better in all cases.

Table 5: Precision, Recall and F-value for all classifiers

Feature Selection	Classifier	Precision	Recall	F-Value
Without FS	FRNN	0.766111277	0.768220223	0.767164301
	FNN	0.795853141	0.807631507	0.801699065
	FRONN	**0.863514768**	**0.862196784**	**0.862855272**
Gain Ratio	FRNN	0.882873532	0.880948487	0.881909959
	FNN	0.879661386	0.88081221	0.880236422
	FRONN	**0.908619183**	**0.90678659**	**0.907701962**
Information Gain	FRNN	0.927028573	0.927609703	0.927319047
	FNN	0.932846953	0.937448896	0.935142263
	FRONN	**0.968523928**	**0.966966476**	**0.967744575**

Table 5 shows values of precision, recall and f-value for all classifiers with and without feature selection. It is noticed that in all cases Fuzzy Rough Ownership NN classifier produces the best performance.

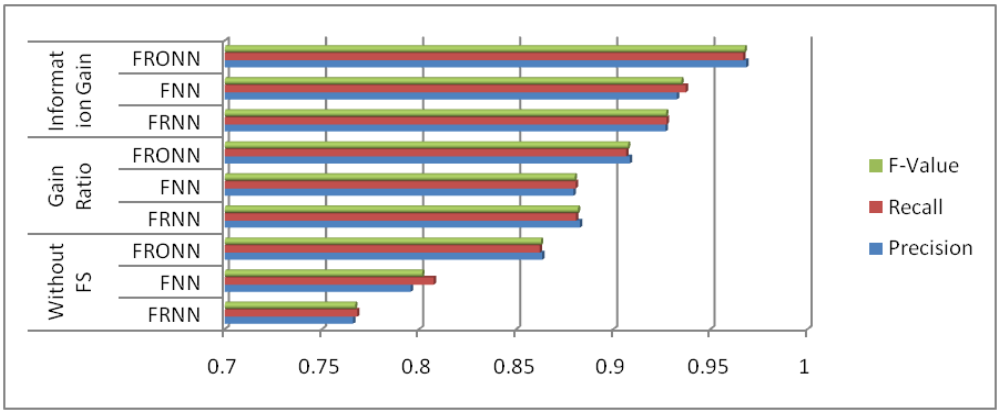


Figure 3: Graphical representation of Precision, Recall and F-value for 3 classifiers

5 Conclusion

To enhance the classification accuracy several approaches have been adopted by researchers. Here, we applied Fuzzy Nearest Neighbor, Fuzzy Rough Nearest Neighbor and Fuzzy Rough Ownership Nearest Neighbor techniques without feature selection. Next, the same techniques are used with feature selections and improvement in classification accuracy is observed. Lastly, weighted average accuracy algorithm is used and further improvement is obtained. In all cases i.e. for accuracy, precision, recall, f-value the Fuzzy Rough Ownership Nearest Neighbor classifier and Information Gain feature selection produced better performance.

ACKNOWLEDGEMENTS:

We would like to thank Dr. E. Al-Masri and Dr. Q.H. Mahmoud for providing us the QWS dataset, which is used in our experiments.

REFERENCE

[1]. LI Yuan-jie and CAO Jian; "Web Service Classification Based on Automatic Semantic Annotation and Ensemble Learning"; 2012 IEEE; DOI 10.1109/IPDPSW.2012.280; pp.2274-2279.

[2]. Venkataiah Vaadaala, R. Rajeswara Rao and Venkateswara Rao .K; " Classification of Web Services Using JForty Eight";International Journal of Electronics Communication and Computer Engineering; ISSN 2249–071X; Volume 4, Issue (6) NCRTCST-2013, pp.181-184.

[3]. Ramakanta Mohanty, V. Ravi and M.R. Patra; "Web-services classification using intelligent techniques"; Elsevier, Expert Systems with Applications 37 (2010); pp. 5484–5490.

[4]. Hongbing Wang, Yanqi Shi, Xuan Zhouy, Qianzhao Zhou, Shizhi Shaoand AthmanBouguettayay; "Web Service Classification using Support Vector Machine"; 2010 IEEE; DOI 10.1109/ICTAI.2010.9; pp.3-6.

[5]. Ramakanta Mohanty, V. Ravi and M. R. Patra, "Classification of Web Services Using Bayesian Network", Journal of Software Engineering and Applications, 2012, 5, 291-296.

[6]. V.Mohan Patro and Manas Ranjan Patra; "Augmenting Weighted Average with Confusion Matrix to Enhance Classification Accuracy"; Transactions on Machine Learning and Artificial Intelligence, UK; ISSN: 2054-7390; DOI: 10.14738/tmlai.24.328; Volume 2 No 4, Aug (2014), pp: 77-91.

[7]. Jesen, R. and Cornelis, C., "A new approach to fuzzy-rough nearest neoghbour classification", LNAI 5306, Springer-Verlag, pp. 310-319(2008).

[8]. Ashalata Panigrahi and Manas Ranjan Patra, "A Hybrid Model for Intrusion Detection Using Fuzzy Rough Theory with Feature Reduction", International Journal of Computer Networks and Security, Vol.23, Issue.2, 1184-1191, Recent Science Publications, ISSN:2051-6878

[9]. http://www.uoguelph.ca/~qmahmoud/qws/dataset/ last accessed on 04/09/14.

[10]. Al-Masri, E., and Mahmoud, Q. H., "Discovering the best web service", (poster) 16th International Conference on World Wide Web (WWW), 2007, pp. 1257-1258.

[11]. Al-Masri, E., and Mahmoud, Q. H., "QoS-based Discovery and Ranking of Web Services", IEEE 16th International Conference on Computer Communications and Networks (ICCCN), 2007, pp. 529-534

[12]. Al-Masri, E., and Mahmoud, Q.H., "Investigating Web Services on the World Wide Web", 17thInternational Conference on World Wide Web(WWW), Beijing, April 2008, pp. 795-804. (for QWS WSDLs Dataset Version 1.0)

[13]. www.cs.waikato.ac.nz/ml/weka/ last accessed on 14/11/14

[14]. Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", 2nd ed., Morgan Kaufmann Publishers, March 2006, ISBN 978-1-55860-901-3.

TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

Offset Phase Shift Keying Modulation in Multiple-Input Multiple-Output Spatial Multiplexing

Adeyemo, Z. Kayode, Rabiu, E. Oluwatosin and Robert, O. Abolade
Department of Electronic and Electrical Engineering
Ladoke Akintola University of Technology, Ogbomoso, Nigeria.
zkadeyemo@lautech.edu.ng

ABSTRACT

The increasing demand for multimedia data transmission in mobile wireless communication poses a challenge to reliable signal reception. In order to have good signal quality, a robust digital modulation scheme is required at the transmitter. However, the conventional M-ary Phase Shift Keying (MPSK) commonly used in multiple-input multiple-output (MIMO) communication systems with non-linear radio frequency (RF) power amplifiers causes a relative increase in intercarrier interference (ICI). This paper presents a development of offset-MPSK (O-MPSK) modulation scheme in MIMO spatial multiplexing over the Rayleigh fading channel.

The O-MPSK modulation schemes were developed for 4, 8, 16, 32 and 64 constellation sizes. The development of the O-MPSK was done by shifting the phase of the conventional QPSK, 8-PSK, 16-PSK, 32-PSK and 64-PSK by an odd multiple of pi (π) to give $\pi/2$-QPSK, $\pi/4$-8PSK, $\pi/8$-16PSK, $\pi/16$-32PSK and $\pi/32$-64PSK, respectively, with a view to reducing the spectral spreading in the power amplifiers at the receiver of a MIMO system. The MIMO techniques used was MIMO Spatial Multiplexing (MIMO–SM). The system models were developed around these schemes and later simulated using MATLAB application toolkit. The performances of the O-MPSK schemes were evaluated using bit error rate (BER) at signal-to-noise ratio (SNR) range of 0 to 20 dB and compared with the conventional MPSK schemes.

The results obtained for all the SNRs in MIMO-SM showed that mean BER of 0.0024, 0.0040, 0.0085, 0.0183 and 0.036 were obtained for $\pi/2$-QPSK, $\pi/4$-8PSK, $\pi/8$-16PSK, $\pi/16$-32PSK and $\pi/32$-64PSK respectively as against mean BER of 0.0025, 0.0044, 0.0088, 0.0178 and 0.0358 obtained for conventional QPSK, 8PSK, 16PSK, 32PSK and 64PSK respectively.

The mean BER values obtained reveal that the developed O-MPSK outperforms the conventional MPSK due to the relatively lower BER of O-MPSK schemes compared with the MPSK schemes. This is as a result of the reduction in the amplitude variations and spectral spreading at the receiver of the MIMO system.

Keywords: Offset, digital modulation, Spatial Multiplexing, MIMO, M-PSK scheme, Intercarrier Interference

1 Introduction

The goal of an ideal digital wireless communication system is to produce the exact replica of transmitted data at the receiver [1]. This has necessitated the corresponding numerous tremendous researches carried out in digital communications industry which leads to rapid growth recorded in

DOI: 10.14738/tnc.32.1144
Publication Date: 28th April, 2015
URL: http://dx.doi.org/10.14738/tnc.32.1144

the past two decades especially in its various applications [2]. This growth, in turn, has spawned an increasing need to seek automated methods of analyzing the performance of digital modulation types using the latest mathematical software or programming language. Digital modulation schemes practically in use now are Amplitude Shift Keying (ASK), Frequency Shift Keying (FSK), Phase Shift Keying (PSK) and Quadrature Amplitude Modulation (QAM) with each having their distinctive features and characteristics. In the case of ASK, the use of amplitude modulated analogue carriers to transport digital information always results in a relatively low quality output. Although it is a low cost type of digital modulation, this is seldom used except for a very low speed telemetry circuits. FSK has a poorer error performance than PSK or QAM and consequently is not used regularly for high-performance digital radio systems [3];[4];[5];[6];[7].

QAM is a modulation scheme in which two schemes (ASK and PSK) are combined to improve the performance of the conventional counterpart modulation making this technique a little complex [8];[9]. It is mainly used for few specific applications [10]. The PSK schemes have constant envelope but discontinuous phase transitions from symbol to symbol and it is the most commonly used digital modulation technique. Some multi-level modulation techniques that permit high data rates within fixed bandwidth known as M-ary PSK schemes are employed in quasi-optical wireless array applications, compressed image communication in mobile fading channel, space applications, Tracking and Data Relay Satellites System (TDRSS) [1].

The demands for high data rate wireless communication in recent years have continued to increase rapidly for wireless multimedia services. Multiple-input, multiple-output (MIMO) systems are now the popular approaches to meet these demands [11];[12]. The use of multiple antennas at both transmitter and receiver in wireless communication links provides a means of maximizing the system performance of wireless systems. MIMO technology provides diversity by making the receiver to receive multiple replicas of the same information-bearing signal; and this provides a more reliable signal reception [13];[14];[15];[16];[17].

The conventional M-ary phase shift keying such as QPSK does not have constant amplitude for transition with a phase shift equal to π ($\pm 180^{o}$). The phase transitions make QPSK signals suffer from large envelope variations when passed through a nonlinear power amplifier operating at saturation. The resulting effects are nonlinear amplitude and phase distortions which cause spectral spreading of the transmitted signal, intercarrier interference (ICI) and degradation of the performance of the communication system. However, Offset Quadrature Phase Shift Keying, OQPSK, suffers from lower envelope variations as a result of the smaller phase transitions as each transition is limited to $\pm 90^{o}$. This results in relatively more constant envelope after pulse shaping [18];[19].

Many researchers have worked using different modulation schemes like [20] carried out performance analysis of a 2x2 spatial multiplexing MIMO technique with high order M-PSK and a combination of ZF and minimum mean square error (MMSE) equalizers without channel state information (CSI) at the transmitter. The system was simulated over the Rician fading channel. Simulation results showed significant improvement in BER performance at SNR value of 40 dB and above for 32-PSK, 64-PSK, 128- PSK, 256-PSK, 512-PSK and 1024-PSK. The implication of the results is that more power is needed to achieving a target BER; and this would not provide the desired power efficiency of the system especially for mobile applications. Mangla and Singh in [22] compared the BER performances of higher order M-QAM and M-PSK modulation schemes in a MIMO-OFDM system. The system was simulated for M = 16, 64, 256, 512 and 1024. The results showed that spectral efficiency increases with increasing modulation order M. Also, M-QAM gives better BER performance than M-PSK. The BER of the higher order modulations can be reduced but at the cost of

increasing the SNR. Increasing the SNR is however not advisable because excessive power consumption would adversely affect system lifespan. Hence, this paper presents O-MPSK in MIMO spatial multiplexing (MIMO-SM) communication systems in order to reduce the ICI towards improving the systems' performances.

2 Development of the Offset M-PSK Schemes

A modulated signal consists of a combination of the carrier signal and the message (or information) signal. The *M*-ary PSK modulation is achieved by shifting the carrier in phase according to the message data. A modulated signal *s(t)* in time (t) domain can be expressed as:

$$s(t) = Re\{\mathbb{g}(t)\exp(j\omega_c t)\} \tag{1}$$

where Re{.} denotes the real component of the complex function indicated by j,
$\omega_c = 2\pi f_c$,
f_c *= the carrier frequency,*
$\mathbb{g}(t)$ *= the complex baseband envelope of* $s(t)$.
This complex baseband envelope $\mathbb{g}(t)$ is a function of the message signal $m(t)$ and can be expressed as:

$$\mathbb{g}(t) = Am(t)\exp[j\theta(t)] \tag{2}$$

where A is a constant amplitude
$\theta(t)$ *= the phase of the signal*
Substituting equation (2) into (1) gives:

$$s(t) = Am(t)\cos[\omega_c t + \theta(t)] \tag{3}$$

Applying trigonometric identity to Equation (3), the equation can be expressed in cosine and sine forms as:

$$s(t) = Am(t)[cos\omega_c t cos\theta(t) - sin\omega_c t sin\theta(t)] \tag{4}$$

The constant amplitude *A* is a function of the signal power; and it is given as:

$$A = \sqrt{2P} \tag{5}$$

where *P* is the signal power. Also, *P* is a function of the energy contained in symbol duration; and is given as:

$$P = \frac{E}{T_s} \tag{6}$$

where T_s *is the symbol period;*
E is the energy contained in the symbol period.
Substituting Equation (6) into (5) gives:

$$A = \sqrt{\frac{2E}{T_s}} \tag{7}$$

With' A ' into equation 4 gives:

$$s(t) = m(t)\sqrt{\frac{2E}{T_s}}[cos\omega_c t cos\theta t - sin\omega_c t sin\theta t] \tag{8}$$

By shifting the carrier in phase, Equation 8 becomes:

$$s(t) = m(t)\sqrt{\frac{2E}{T_s}}[cos\omega_c t cos(\theta_i - \theta_0)t - sin\omega_c t sin(\theta_i - \theta_0)t] \tag{9}$$

with

$$\theta_i = \frac{2\pi}{M} i, \quad for\ i = 1,2,3,\ldots M \tag{10}$$

and θ_o is the initial phase given as:

$$\theta_0 = \frac{2\pi}{M} \tag{11}$$

where *M* is the constellation size of the *M*-ary PSK; the phase takes on one of *M* possible values.

Equation (9) represents an *M*-ary PSK modulated signal. The phases of an MPSK constellation can be represented with a polar diagram in Inphase/Quadrature (I/Q) format. The cosine component of the modulated signal $s(t)$ takes the inphase axis while the sine component takes the quadrature axis.

The offset MPSK (OMPSK) modulation can be implemented by delaying the input bit stream of the quadrature part by one bit period T_b. The bit period is given as:

$$T_b = \frac{T_s}{k} = \frac{T_s}{\log_2 M} \tag{12}$$

where k is the number of bits that represents a symbol. Therefore, the conventional MPSK modulation Equation 9 can be modified for the OMPSK as:

$$s(t) = m(t)\sqrt{\frac{2E}{T_s}}\left[\cos\omega_c t\cos(\theta_i - \theta_0)t - \sin\omega_c t\sin[(\theta_i - \theta_0)(t - T_b)]\right] \tag{13}$$

2.1 Offset QPSK Scheme

The least complex form of the offset *M*-ary PSK is the offset 4-ary PSK (OQPSK). For the OQPSK the number of bits per symbol k is 2; hence, the bit period $T_b = T_s/2$. Also, Equation 11 shows that the phase is shifted by $\pi/2$ when *M* = 4. Figure 1(a) shows the designed scheme for implementing OQPSK modulation. The modulation is achieved by transmitting the odd-numbered input bits via the inphase, I(t) branch while the even-numbered bits are transmitted via the quadrature, Q(t) branch with the use of a serial-to-parallel (S/P) converter. The data on the Q(t) part is delayed by $T_s/2$ with respect to that on the I(t) part to create an offset. This is followed by unipolar-to-bipolar (U/B) converters which convert the binary data to polar non return-to-zero (NRZ). The bipolar (±1) signals are then passed through rectangular pulse-shaping filters, and then modulated by cosine and sine carriers.

At the receiver, the signal is demodulated as shown in Figure 1(b). The arriving OQPSK signal is passed through a carrier recovery circuit which demodulates the signal by the inphase and quadrature carriers. The resulting I(t) and Q(t) signals are then passed through the Integrate and Dump filters followed by the detection of the binary data by the threshold detector. The inphase, I(t) stream is then delayed with respect to that on the Q(t) stream by $T_s/2$ to remove the offset introduced at the modulator. The I(t) and Q(t) streams are then combined by a parallel-to-serial (P/S) converter to produce the received bit stream. This strategy helps to reduce spectral spreading when the signal passes through a nonlinear high power amplifier because the offset makes the signal to have lower envelope variation when compared with the conventional QPSK.

2.2 Offset 8-PSK Scheme

The offset 8-PSK (O8-PSK) modulator and demodulator are shown in Figures 2(a) and (b) respectively. The modulation and demodulation processes of O8-PSK are similar to those of OQPSK except that the phase is shifted by $\pi/4$ and the number of bits per symbol k is 3.

2.3 Offset 16-PSK Scheme

The offset 16-PSK (O16-PSK) modulator and demodulator are shown in Figures 3(a) and (b) respectively. The modulation and demodulation processes of O16-PSK are similar to those of OQPSK except that the phase is shifted by $\pi/8$ and the number of bits per symbol k is 4.

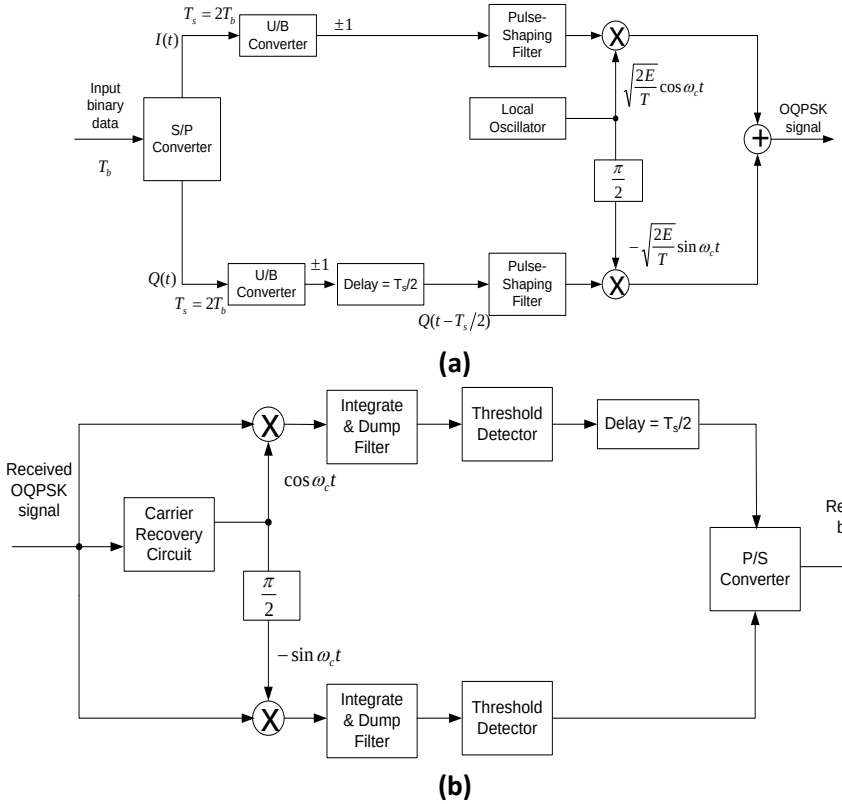
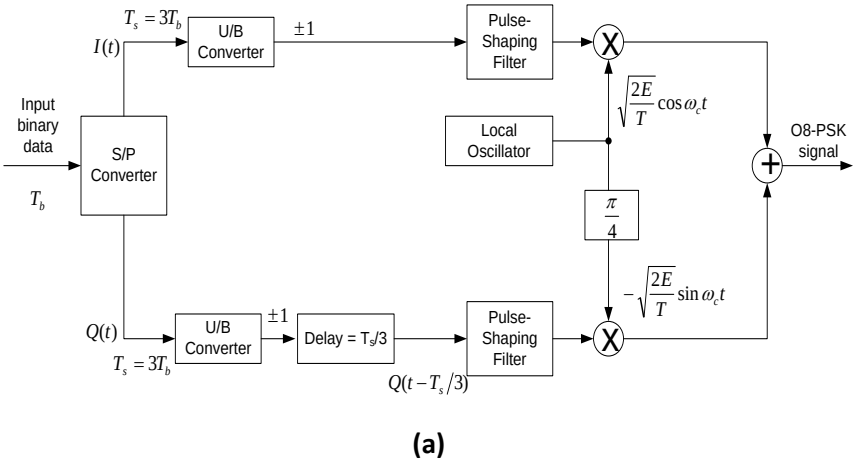
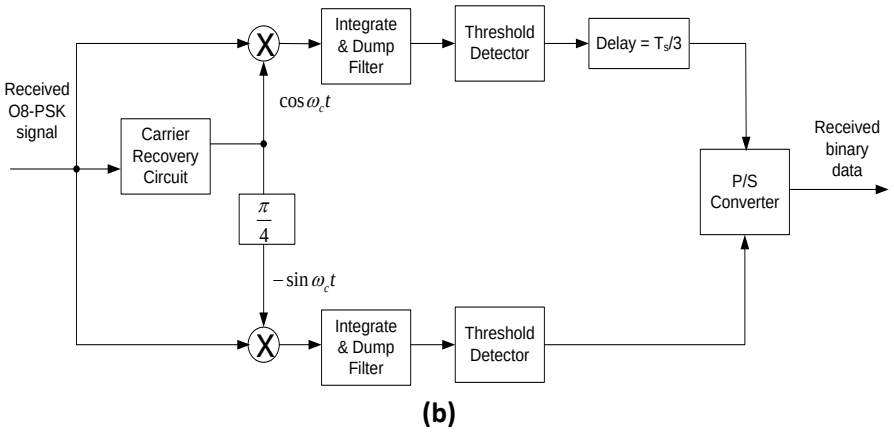


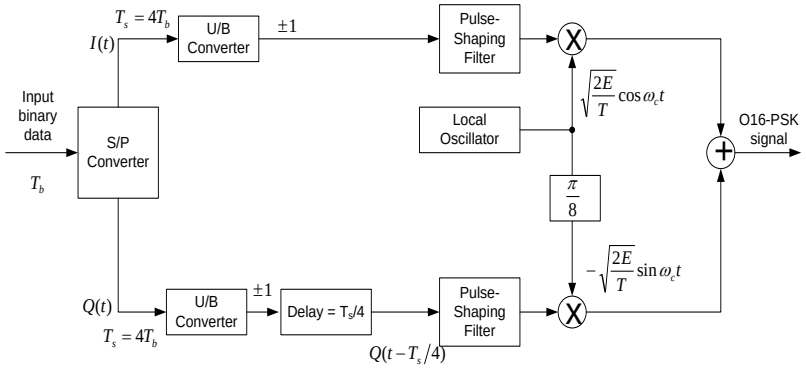
Figure 1: Offset QPSK Scheme (a) Modulator (b) Demodulator



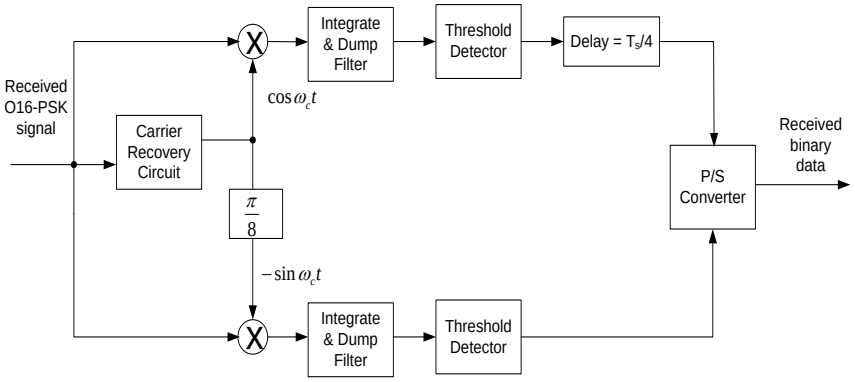


(b)

Figure 2: Offset 8-PSK Scheme (a) Modulator (b) Demodulator



(a)



(b)

Figure 3: Offset 16-PSK Scheme (a) Modulator (b) Demodulator

2.4 Offset 32-PSK Scheme

The offset 32-PSK (O32-PSK) modulator and demodulator are similar to those of OQPSK except that the phase is shifted by $\pi/16$ and the number of bits per symbol k is 5.

2.5 Offset 64-PSK Scheme

The modulation and demodulation processes of O64-PSK scheme are similar to those of OQPSK except that the phase is shifted by $\pi/32$ and the number of bits per symbol k is 6.

3 System Simulation Model and Tool

The investigation was carried out by developing the MIMO-SM system simulation models using the OMPSK modulation schemes. The developed system models were implemented by simulation of the developed models.

3.1 System Simulation Models

A simulation model for 2x2 MIMO-SM with OMPSK modulation scheme is shown in Figure 4. The transmitted message is a randomly generated bit stream. At the transmitter, the bit stream is passed through the OMPSK modulator. The output signal from the modulator is split into even and odd symbols; and the even symbols are transmitted through the antenna 1 while the odd symbols are transmitted through antenna 2. The two signals pass through a Rayleigh fading channel with the additive white Gaussian noise (AWGN). At the receiver, channel estimation is performed on the received signals to nullify the effect of fading. The signals are then multiplexed and the resulting OMPSK signal is demodulated to obtain the received binary data.

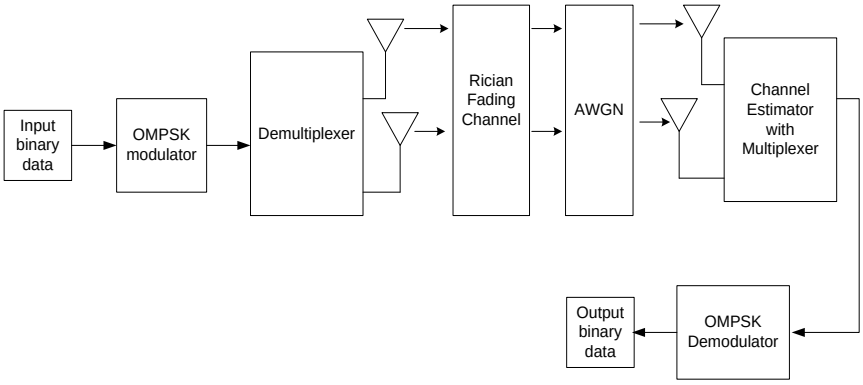


Figure 4: System simulation model for 2x2 MIMO-SM with OMPSK modulation scheme

4 Results and Discussion

MIMO Spatial Multiplexing (MIMO-SM) and MIMO Beamforming (MIMO-BF) schemes were simulated using the O-MPSK modulation technique as well as the conventional MPSK modulation technique over a Rayleigh fading channel; and comparisons are made in terms of BER between the two schemes. The BER is evaluated for SNR values of 0 to 20 dB as presented in Table 1. The BER performances of QPSK and $\pi/2$-QPSK in MIMO-SM over Rayleigh fading channel are shown in Figure 5. Taking SNR of 10 dB, QPSK gives a BER value of 0.0012 while $\pi/2$-QPSK gives 0.0010; also, the mean BER values for all the SNRs are 0.0025 and 0.0024 for QPSK and $\pi/2$-QPSK respectively. The subsequent lower BER values given by $\pi/2$-QPSK reveals the efficiency of $\pi/2$-QPSK over the QPSK.

Figure 6 presents the BER performances of 8PSK and $\pi/4$-8PSK in MIMO-SM over Rayleigh fading channel. Taking SNR of 10 dB, the BER values for 8PSK and $\pi/4$-8PSK are 0.0018 and 0.0016 respectively, and the mean BER values are 0.0044 and 0.0040 for 8PSK and $\pi/4$-8PSK respectively. The $\pi/4$-8PSK has relatively lower BER compared to 8PSK. The BER performances of 16PSK and $\pi/8$-16PSK in MIMO-SM over Rayleigh fading channel are shown in Figure 7; the 16PSK modulation gives

a BER value of 0.0044 while the $\pi/8$-16PSK modulation gives a closer 0.0045 at SNR of 10 dB. The mean BER values for 16PSK and $\pi/8$-16PSK are 0.0088 and 0.0085 respectively. This result reveals that the $\pi/8$-16PSK has relatively better BER performance compared to 16PSK.

Table 1: BER values for 2x2 MIMO spatial multiplexing over Rayleigh fading channel

SNR[dB]	QPSK	$(\frac{\pi}{2})$-QPSK	8PSK	$(\frac{\pi}{4})$-8PSK	16PSK	$(\frac{\pi}{8})$-16PSK	32PSK	$(\frac{\pi}{16})$-32PSK	64PSK	$(\frac{\pi}{32})$-64PSK
0	0.0075	0.0083	0.0153	0.0136	0.0310	0.0279	0.0585	0.0583	0.1040	0.1043
2	0.0064	0.0062	0.0117	0.0110	0.0215	0.0208	0.0430	0.0444	0.0796	0.0807
4	0.0045	0.0044	0.0084	0.0078	0.0158	0.0156	0.0323	0.0336	0.0634	0.0642
6	0.0032	0.0029	0.0056	0.0053	0.0104	0.0109	0.0221	0.0235	0.0470	0.0476
8	0.0022	0.0017	0.0032	0.0031	0.0071	0.0074	0.0153	0.0157	0.0344	0.0349
10	0.0012	0.0010	0.0018	0.0016	0.0044	0.0045	0.0100	0.0105	0.0244	0.0245
12	0.0009	0.0007	0.0011	0.0009	0.0026	0.0026	0.0065	0.0069	0.0170	0.0172
14	0.0006	0.0005	0.0005	0.0005	0.0017	0.0017	0.0040	0.0041	0.0110	0.0108
16	0.0004	0.0003	0.0003	0.0004	0.0010	0.0008	0.0026	0.0027	0.0070	0.0068
18	0.0002	0.0002	0.0002	0.0002	0.0005	0.0005	0.0011	0.0012	0.0042	0.0038
20	0.0002	0.0001	0	0	0.0004	0.0004	0.0007	0.0006	0.0014	0.0013

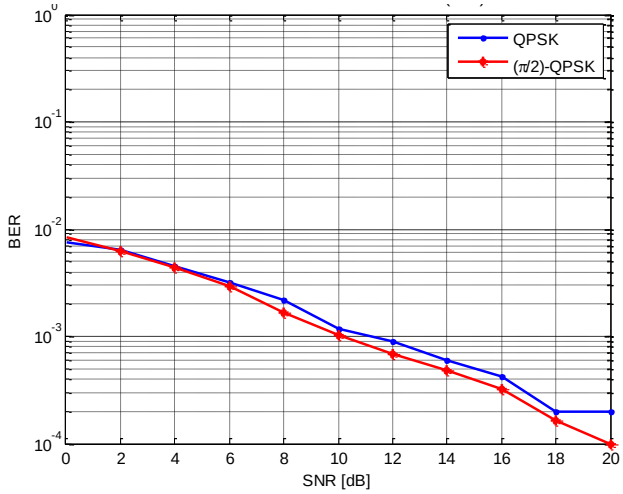


Figure 5: BER performances of QPSK and $\pi/2$-QPSK in2x2 MIMO-SM over Rayleigh Fading Channel

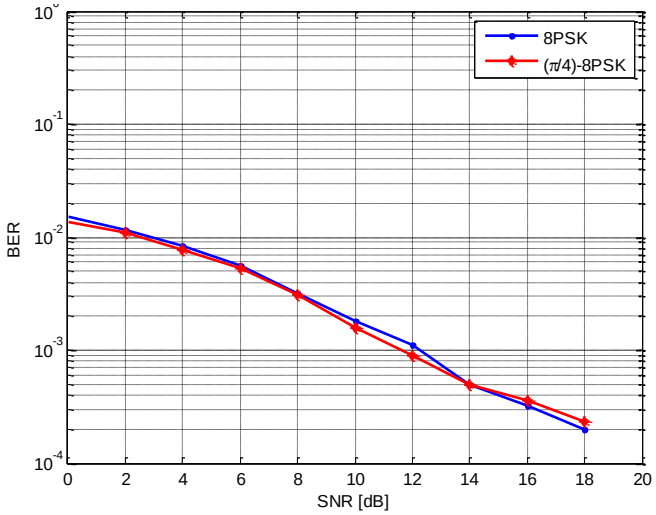


Figure 6: BER performances of 8PSK and $\pi/4$-8PSK in2x2 MIMO-SM over Rayleigh Fading Channel

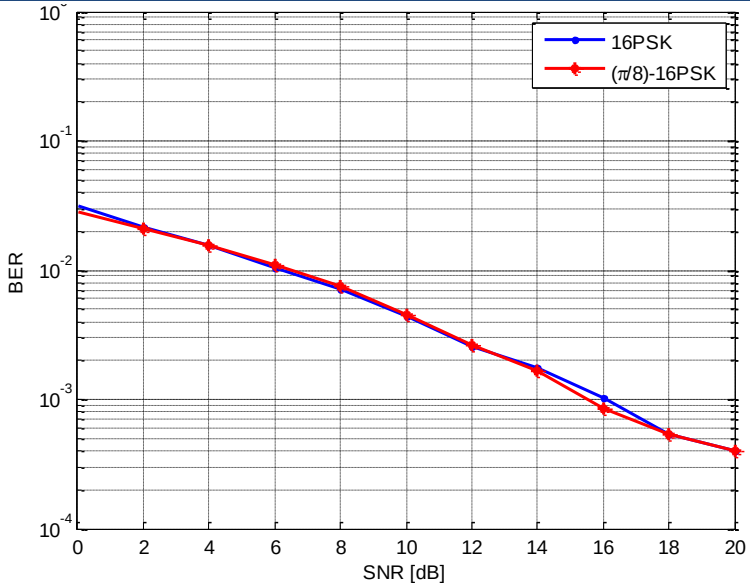


Figure 7: BER performances of 16PSK and $\pi/8$-16PSK in2x2 MIMO-SM over Rayleigh Fading Channel

The BER performances of the off-set M-PSK in MIMO-SM for comparison over Rayleigh fading channel are presented in Figure 8. At SNR of 10 dB, the BER values for 32PSK and $\pi/16$-32PSK are 0.0100 and 0.0105 respectively, and the mean BER values are 0.0178 and 0.0183 for 32PSK and $\pi/16$-32PSK respectively. This result revealed that 32PSK has better BER performance compared to $\pi/16$-32PSK. The BER performances of 64PSK and $\pi/32$-64PSK gave BER values of 0.0244 and 0.0245 respectively at SNR of 10 dB. The mean BER values for 64PSK and $\pi/32$-64PSK are 0.0358 and 0.036 respectively. This result revealed that the 64PSK gave relatively better BER performance compared to $\pi/32$-64PSK.

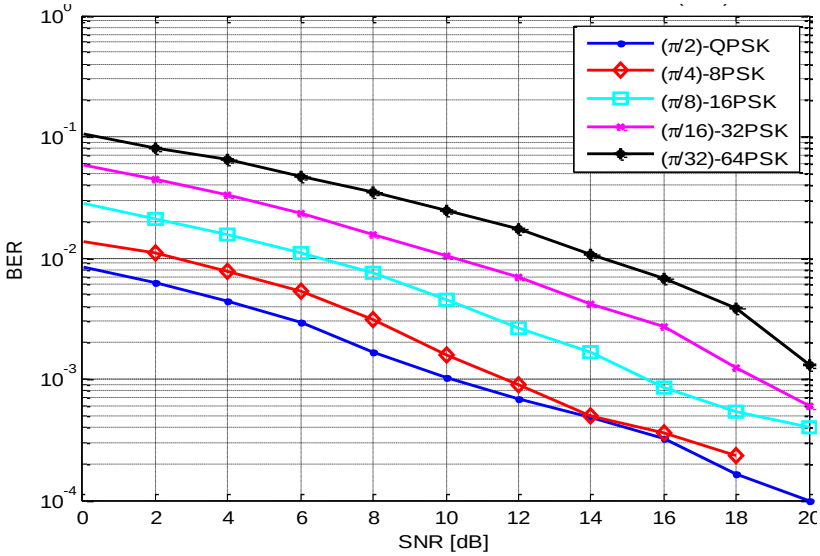


Figure 8: Comparison of the O-MPSK in 2x2MIMO-SM over Rayleigh Fading Channel

5 Conclusion

In this paper, O-QPSK, O-8PSK, O-16PSK, O-32PSK and O-64PSK modulation schemes have been developed, the system simulation models incorporating a 2x2 MIMO Spatial Multiplexing (MIMO–SM) using the O-MPSK modulation schemes was developed over Rayleigh fading channel, simulated

using MATLAB application package. The models were evaluated to determine the performance using bit error rate (BER) and compare with the conventional (non-offset) MPSK. The O-MPSK schemes were compared with MPSK scheme in terms of BER for SNR values of 0 to 20 dB. The O-MPSK investigated include $\pi/2$-QPSK, $\pi/4$-8PSK, $\pi/8$-16PSK, $\pi/16$-32PSK and $\pi/32$-64PSK. The results revealed that the O-MPSK schemes outperform the MPSK schemes in MIMO-SM as the O-MPSK schemes gave relatively lower mean BER compared to the MPSK schemes. Also, the results revealed that the best performance was obtained with the $\pi/2$-QPSK scheme.

REFERENCES

[1]. Harjot K., Bindiya J. and Amit V., "Comparative Performance Analysis of M-ary PSK Modulation Schemes using Simulink" International Journal of electronic & Communication Technology, 2. 3, (2011): 204 – 209.

[2]. Gesse, M. K. and Oladele O. P., "Performance Evaluation of LTE Downlink with MIMO Techniques," M.Tech Thesis, Blekinge Institute of Technology, Karlskrona, Sweden, (2010).

[3]. Proakis, J. G. and Salehi, M., Communication Systems Engineering, 2nd Ed., Prentice-Hall Inc., USA, (2002).

[4]. Proakis, J. G., Digital Communications, McGraw – Hill Companies, inc, International Edition, New York City, USA, (2001).

[5]. Rappaport, T. S., Wireless Communications – Principles and Practice, 2nd Ed., Prentice-Hall Inc., USA, (2002).

[6]. Bernard, S., "Digital Communications: Fundamentals and Applications", 2nd edition, Prentice Hall, USA, (2001).

[7]. Amin, A. (2011), "Computation of Bit-Error Rate of Coherent and Non-Coherent Detection M-ary PSK with Gray Code in BFWA Systems," International Journal of Advancements in Computing Technology, 3.1 (2011):27-38.

[8]. Antti V. R. and Arto L., "Radio Engineering for Wireless Communication and Sensor Applications", Artech House, Boston, London, (2003)

[9]. Poongodi, C., Ramya, P. and Shanmugam A., "BER Analysis of MIMO OFDM System using M-QAM over Rayleigh Fading Channel," International Conference on Communication and Computational Intelligence, Kongu Engineering College, Perundurai, erode, T. N., India, (2010): 284-288.

[10]. Ippolito, L. J. "Satellite communication system engineering: Atmospheric Effect, Satellite link Design, and System Performance," first edition John Wiley & Sons, Singapore, (2008).

[11]. Foschini G. J. and Gans M., "On Limits of Wireless Communications in a Fading Environment when using Multiple Antennas," Wireless Personal Communications, 6,(1998):311-335,

[12]. Sulyman, A. I., "Performance of MIMO Systems with Antenna Selection over Nonlinear Fading Channels," IEEE Journal of Selected Areas in Communications, 2.2, (2008):159 – 170.

[13]. Gesbert D., Bolcskei H., Gore D. and Paulraj A., "MIMO Wireless channels: Capacity and Performance Prediction," in Proc. IEEE Globecom, San Francisco, CA, (2000):1083 - 1088

[14]. Almouti, S. M., "A Simple Transmit Diversity Technique for Wireless Communications," IEEE Journal on Select Areas in Communications, 16.8, (1998):1451-1458.

[15]. Mindaudu, A. S. and Miyim, A. M., "BER Performance of MPSK and MQAM in 2x2 Almouti MIMO Systems," International Journal of Information Sciences & Techniques, 2.5, (2012):1 – 10.

[16]. Jaiswal A. K., Kumar A. and Singh A. P., "Performance Analysis of MIMO-OFDM System in Rayleigh Fading Channel," International Journal of Scientific and Research Publications, 2. 5, (2012): 1 – 5.

[17]. Kaur N. and Kansal L., "Performance Comparison of MIMO Systems over AWGN and Rician Channels with Zero Forcing Receivers," International Journal of Wireless and Mobile Networks, 5.1, (2013):73 – 84.

[18]. Nelson, T. and Rice, M., "MIMO Communications using offset modulations," Proceedings of IEEE International Waveform Diversity and Design Conference, (2006): 23 – 27.

[19]. Dang, X., Offset QPSK in SISO and MIMO Environments, PhD Dissertation, Brigham Young University, Provo, Utah, (2009).

[20]. Vadhera A. and Kansal L., "BER Analysis of 2x2 MIMO Spatial Multiplexing under AWGN and Rician Channels for Different Modulations Techniques," International Journal of Wireless and Mobile Networks, 5.5, (2013):85 – 98.

[21]. Mangla R. and Singh M., "Performance Comparison of MIMO-OFDM Transceiver Wireless Communication System using QAM and QPSK Modulation Schemes," International Journal of Advances in Engineering Science and Technology, 1.2, (2012): 66 – 72.

TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

Developing of Human Resources in E-learning and Practical Experience in its implementation

[1]Tamar Gogoladze, [2]Natia Zhozhuashvili, [3]Elene Khojevanishvili and [4]Ana Tsiklauri
[1]tamilagogoladze@gmail.com; [2]natia7413@gmail.com; [3]ekhojevanishvili@gmail.com;
[4]ana.tsiklauri@gmail.com

ABSTRACT

The article deals with practical experience in developing and implementing human resources in E-learning in Gori State teaching University.

German International cooperation civil organization (GIZ) contributed greatly the development of electronic learning sphere during a 7-year cooperation with Georgia. The working groups sent from Gori State teaching University after completion the full course of trainings and getting certificates initiated implementing the project "The Development of Human Recourses for implementation of E-learning" and planned to hold various activities within the project, also the research for finding out the readiness for implementing the innovation was held. The results got by the feedback clarified that the professors and teachers of the University share the necessity of implementing e-learning in the teaching process as an assisting tool and express their readiness for preparation pilot programs. After competition of the full training course (4 modules) the y presented presentations in 10 different disciplines which were greatly approved by GIZ experts. The project participants and leaders made further plans for implementation and development of E-learning.

For the country like Georgia, it is important to raise the competences and skills of professors and teachers and mobilize all the resources to meet the challenges of the epoch.

1 Introduction

The availability of computer and internet access in the XXI century enabled modern methodology to use online learning widely. E-learning system is a good example of this, using information and electronic technologies where "multimedia design" plays an important role.

According to R.Mayer "multimedia design" enables perception of material by a human when it is presented by texts and illustrations, it is concentrated on continuity, it is modal, it has outlined structure (with titles and sub-subtitles). Besides, it should also be taken into consideration that such learning should meet student's needs, expectations, possibilities, it should develop the skills and competences that will answer modern challenges, reduce expenses and maintain personal freedom. Besides the applicants for studying at universities it should also meet the Andragogy needs that is- it should be available to the people seeking job or attempting to get a higher position. According to one of the strategies e-learning is considered as an instrument based on knowledge, for creating dynamic, competitive economics and creating lifelong learning space.

The world industry of E-learning is about 48 milliard dollars. It was created as a result of developing internet and multimedia. Electronic learning is the usage of communicative technologies and internet in teaching process as an assisting instrument for filling and not substituting traditional

DOI: 10.14738/tnc.32.1087
Publication Date: 28th April, 2015
URL: http://dx.doi.org/10.14738/tnc.32.1087

learning. The implementation of the e-learning component aims rising teaching quality and increasing its affectivity. Its key moments are consulting, content, technologies, service and support and it should be taken into consideration that "virtual model"appeared to be more competitive than other forms of teaching and became a main tool in modernizing of education in many countries of the world and among them it aroused interest in Georgia as well. Nowadays teachers, program providers and teaching developers are the interested part who realizes that exchange of information by means of communicative net is very important for affective teaching.

2 Implementation

German International cooperation civil organization (GIZ) contributed greatly the development of electronic learning sphere during the 6-year cooperation with Georgia. The working groups sent from almost every important Universities of Georgia, separate ministries, various organizations and vocational schools attended general practical course offered by CIZ and as a result of those trainings several (more than ten) pilot programs were created.

A group of professors sent from Gori State Teaching University –T. Gogoladze, E. Khojevanishvili, N. Zhozhuashvili and the main specialist of Information technologies support service-A. Tsiklauri-participated in the training program "The Development of Human Resources and Institutional Building for Electronic Learning in the South Caucasus" organized by GIZ (Elke Vemhoff-Project Manager, Irakli Shurghaia - a coordinator in Georgia). Our involvement in acquiring e-learning principles was conditioned by general and concrete circumstances existed in our higher educational institution.

The training organized by GIZ was held in two parts-in face to face and online regimes. The program consisted of four modules (e-learning management in practice, didactic methodology/instructive design, creating teaching materials and mentor), which were led by certified trainers trained by GIZ. At the training a few factors sprang out, namely, e-learning is a constituent part of Bologna process and a form of student oriented teaching, its motto is "learn whatever you wish, whenever you wish and wherever you wish". Also it was mentioned that leading Universities of Europe and USA have been using e-learning for a long time as an assistant tool and a lively alternative of a traditional learning process and that is why, nowadays the majority of Universities of the world use this method successfully. In France, Finland, South Korea, USA, Great Britain, New Zeeland, Canada and other countries e-learning became practice of everyday educational and labor affairs.

Also the priorities of electronic learning were outlined: free option of time and space, flexible learning conditions, possibility of working in asynchrony format; diversity of materials-video, audio, multimedia or the usage of some other teaching complex elements enables us to make academic process more effective, creative and inclusive, substitution of teaching forms make us avoid routine and increases student's interest and motivation; Constant complement-renovation of materials and possibility of making needy changes enables us 1. To use many times once created material 2. To meet the demandsE-learning course. It gives possibility to use the knowledge and experience of highly qualified personnel working in other companies. It is easy to use for training personnel working in different types of organizations. It is advantageous for continuous education. The monitoring of learning process is simplified and students' testing and evaluation system is perfected. It requires less financial needs. Once created course can be used during a long period simultaneously with several target groups.

On the basis of the information acquired at the trainings and taking into consideration the characteristics of e-learning our group presented a piloted course for the summing up meeting of the training: "Modern Georgian Language" which became a subject of interest and was worthy of the experts attention and after completion of the full course of the training the participants were awarded certificates of electronic learning specialists.

The training of GIZ attended by our group created and idea and motivation to prepare a project " The Development of Human Recourses for Implementation Electronic Learning" at Gory State teaching University with the assistance of GIZ trainers and relying on their experience. This project was carried out in May-June 2014 with active support of the Rector (Prof. G. Sosiashvili). (The Project manager - N. Zhozhuashvili, trainers-T.Gogoladze, E. Khojevanishvili and A. Tsiklauri.) For carrying out the project the following activities were held:

For the administration and professors of the University were held: 1. Presentation-summing "Electronic learning: goals and objectives", 2. Presentation- "The necessity of implementation of E-learning in Higher educational institutions", 3. Research for finding out the readiness for implementation of innovational learning, 4. Training-"Electronic learning: goals, objectives, implementation of innovational learning at educational institutions and getting to know with successful e-learning courses. (Professor Mariam Manjgaladze, GIZ e-learning center expert participated with the project working team).

After the mentioned work with the feedback it became clear that the usage of e-learning as an assistant tool of traditional education, will be acceptable for the participant professors and teachers and for this reason preparing electronic courses will be affective.

The answers to the following questions in percent are presented on the diagram 1.

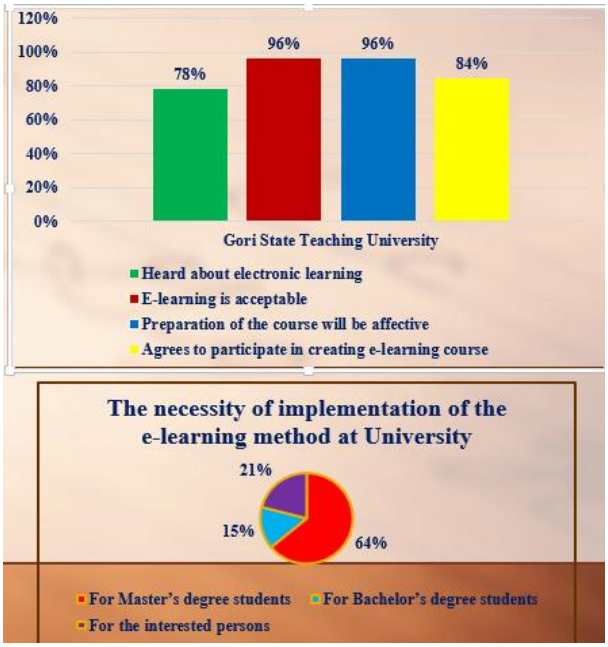


Figure 1

- Have you heard about electronic learning?
- At what rate will be acceptable for you the usage of e-learning as an assistant tool of the traditional education?
- What is your opinion –how affective will be preparing electronic courses in your subject?
- Do you agree your participation in creating electronic course?

- In what perspective do you see the necessity of implementation of e-learning method at University?

That means that the research proved the readiness for preparing human resources for implementation of electronic learning.

The project aimed to prepare teachers and professors for creating electronic courses that on the next stage by implementing e-learning as an assistant tool of traditional academic process facilitate master's degree students and bachelor's degree students in academic process, to create for them more flexible learning environment with less expenses, with online communication to their mentors for consultations. In case of employment or some other extreme situations they will be able to acquire the planned material foreseen by the curriculum. They will be given possibility to prepare materials and homework, acquire knowledge and to get to know with scientific literature related to the given subject in the place and time available for them. By electronic courses they will be provided with materials enriched by more and more and diverse visual, audio-video materials. They will be able to control their own evaluation and finally on the basis of the gained knowledge pass final exams successfully and use the gained skills in their practical work.

The working schedule according to the modules was worked out:

1. Teaching methodology- the goals and objectives of teaching methodology (Preparation); objectives to be solved in the process of implementation and their correlation (implementation); evaluation of the courses and goals. Mentor and tutor; (T. Gogoladze)

2. Creating content (teaching materials) – the stages of implementation of teaching technologies and their meaning in educational institutions; the stages of creating content; (methodological viewpoints, target groups and work with them, reflexion connected to the teaching material, screen segmentation design, instruments for creating content); forming the structure and conception of the content- schematic frame in author program, the draft of schematic frame for programming, gathering material, technical issues; realization –creating teaching materials, interactive elements, content forming evaluation, packing, quality assurance, general instructions; interaction; self-evaluation-feedback; (E. Khojevanishvili)

3. Electronic learning Management – preparation: defining the goals and objectives of the project, forming the standard of the quality; (forming project team-distributing resources, preparing budget; time management-planning process and time; self-checking test); management in practice – implementing; (control of goals, standards and budget; summing up test; following the settings of time); management in practice-packing; (measuring goal achievement; review of time management; evaluation of the usage of resources; calculating real expenses; preparation the next course; self-evaluation test; (NatiaZhozhuashvili)

4. Instruction design review of Caucasus-learning, the essence of Exe; creating design; questionnaire methods. (Exe program), self-evaluation methods (Exe program), uploading video-audio materials (Exe program), glossary. (Ana Tsiklauri)

After carrying out the project the me members of the target team, the professors of various subjects presented prepared points and issues from the pilot electronic courses in History, in various subjects of Economy, Dialect study, Eco-tourism, Computing skills and Academic writing which were highly evaluated by specially invited International trainers –Maia Ninidze and Mariam Manjgaladze. They denoted that within the project great interest and readiness from Teachers and professors of the

University for implementation innovated teaching methods was quite vivid and the result-the goal of the project-was impressive which clarifies once more the possibilities of online cooperation.

The pilot electronic courses presented within the project, in which Information-communication resources, texts, video-audio files, multimedia presentations, teaching websites/blogs were used, proved that the gained knowledge and skills will enable the members-the academic staff of the target team to implement electronic courses and e-learning in the nearest future. In the context of training the academic staff of the University in the sphere of e-learning, on the basis of the experience of other universities the courses on the topics such as individual online consultations, creating electronic dossiers and renewing, testing by means of net tools and others, will be implemented that will find positive reflection on the academic process. Creating e-learning center at the university, development of the infrastructure, technical and program maintenance will be inevitable. Diagram N2:

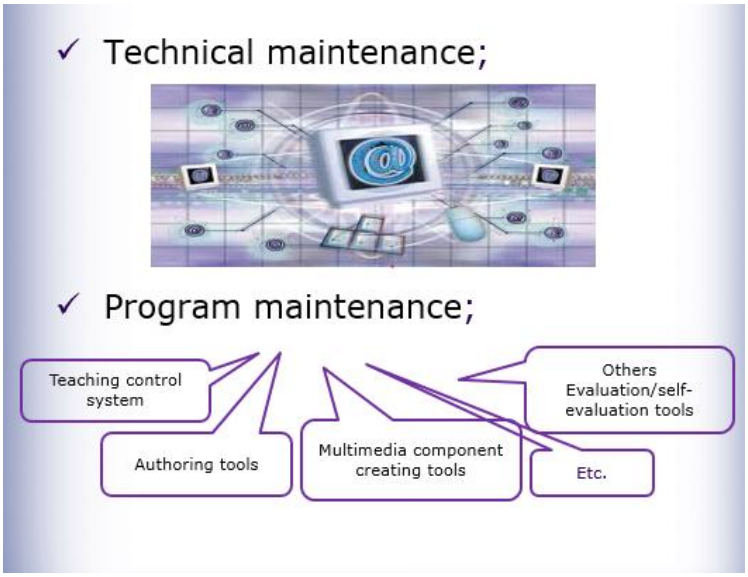


Figure 2

Nowadays, when the Ministry of education and Science is planning to widen teaching sphere by new media and cover the education system, the small country like Georgia, which is facing new challenges inevitably needs rising professors and teachers competences and mobilization of all the resources in order to meet the needs of the epoch honorably.

REFERENCES

[1] Mayer R.E. Multimedia Learning. Cambridge University Press. 2001.

[2] Mayer R. E., Moreno R. Nine ways to reduce cognitive load in multimedia learning, Educational psychologist. 38 (1), 2003.

[3] Najjar L. J. Multimedia information and learning, Journal of educational multimedia and hypermedia. 5(2), 1996.

[4] Сатунина А. Е., Информационное обеспечение стратегических задач в области высшего образования с использованием ЭВМ , Передовой научный опыт высшей школы, рекомендуемый для внедрения. М., 1991. Вып. 2.

[5] Сатунина А. Е., Современное состояние и развитие правовой информатизации высшей школы. М., 1999.

[6] Сатунина А. Е., Сысоева Л. А., Финансы и статистика, Управление проектом корпоративной информационной системы предприятия. М., 2009 .

[7] Полат Е.С., Бухаркина М.Ю., Моисеева М.В., Теориа и практика дистанционного обучения, М., 2004

TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

Membership Protocols for the iTrust Network

[1]Yung-Ting Chuang, [2]Peter M. Melliar-Smith, [3]Louise E. Moser and [4]Isai Michel Lombera
Department of Electrical and Computer Engineering, University of California, Santa Barbara
Santa Barbara, USA
[1]ytchuang@ece.ucsb.edu, [2]pmms@ece.ucsb.edu, [3]moser@ece.ucsb.edu, [4]imichel@ece.ucsb.edu

ABSTRACT

The iTrust system is a decentralized and distributed system for information publication, search and retrieval over the Internet, which is designed to make it difficult to censor or filter information. In this paper, we present four membership protocols for the iTrust network, namely, the non-adaptive, retry, adaptive and combined adaptive membership protocols. We compare the performance of these membership protocols, with respect to four performance metrics, namely, membership accuracy, match probability, response time and message cost, for various parameter values when the membership churn is high and when the membership is stable.

Keywords: Membership protocol, Membership churn, Information publication, search and retrieval, Peer-to-peer network

1 Introduction

Our modern society depends on uncensored publication and retrieval of information over the Internet. Currently, for reasons of efficiency and economy of scale, centralized search engines dominate Internet search and retrieval. However, centralized search engines can easily censor, filter or bias the information they provide. An effective decentralized and distributed search system, as an alternative to conventional centralized search, can help to ensure the free flow of information over the Internet.

The iTrust system [5, 15, 16] is a decentralized and distributed system for information publication, search and retrieval over the Internet, which makes it difficult to censor or filter information. iTrust is based on a peer-to-peer network that is deliberately unstructured to reduce the risk of censorship. The communication cost for iTrust is greater than that for centralized search engines or structured peer-to-peer systems. However, if users suspect that information important to them is being censored or suppressed, they should be willing to incur that extra cost. Nevertheless, we try to minimize any additional cost.

In iTrust, a source node with information to share distributes metadata and a URL for that information, to a subset of the nodes in the membership of the iTrust network, chosen at random. A requesting node, seeking information, generates and distributes its query to a subset of the nodes in the membership, chosen at random. Nodes, that receive metadata and a matching query, send the URL for the information to the requesting node, which retrieves the information from the source node. If the metadata and the requests are distributed to enough nodes, the probability of a match and the consequent retrieval of information are very high [15].

The iTrust system is designed to protect against malicious nodes that attempt to disrupt the search for information. A potential attack might insert, into the membership of the iTrust network, nodes

DOI: 10.14738/tnc.32.1166
Publication Date: 30th April, 2015
URL: http://dx.doi.org/10.14738/tnc.32.1166

that behave normally except that, for sensitive information, they do not match metadata and requests, thus reducing the probability of a match. To guard against such an attack, each node computes, locally and independently, an estimate of the proportion of non-matching nodes in the membership. The node then increases the number of nodes to which it distributes metadata and requests, to maintain the desired high probability of a match.

Another potential attack on iTrust might distribute metadata and URLs for misleading information. iTrust itself does not attempt to detect or downrank such information. If it were to do so, iTrust itself would be engaging in a form of censorship. Rather, as described in [18], a requesting node downloads, from the source node, a table of frequencies of terms in the document. The requesting node matches those terms against its query to generate a relevance ranking for the document. Whether the document is useful or misleading is determined by the user.

iTrust does not attempt to provide anonymity for its users. Anonymity is quite distinct from censorship. Anonymity attempts to hide the identities of users who publish, or search for, information. Censorship attempts to hide the information itself. iTrust could be coupled with an anonymity service [8, 20], so that users could publish or retrieve information without disclosing their identities. We are investigating combining one of those anonymity services with iTrust.

An extensive literature on membership exists, but most of that work is not relevant to iTrust. Previous membership protocols [3] aim to achieve an agreed accurate membership based on a consensus algorithm [2]. The membership protocols for iTrust are simpler and less costly than previous membership protocols, because iTrust does not need to achieve an agreed accurate membership but can operate effectively with an approximate membership for a rapidly changing network. As large-scale peer-to-peer networks become more common, effective approximate membership protocols will become more important.

In iTrust, the membership consists of the nodes that participate in the iTrust network (also referred to as the participating nodes). Each node has a local view of the membership, which the membership protocols aim to keep close to the actual membership. A node can join the membership at any time; likewise, a node can leave the membership at any time, either voluntarily or because it is faulty or disconnected.

In [5, 15, 16], we presented an overview of the iTrust system and the principles on which iTrust is based. In [1], we presented a version of iTrust in which nodes maintain a randomized subset of the membership, a neighborhood. In large networks, containing millions of nodes, the cost of maintaining the entire membership is excessive. iTrust can operate effectively with the membership protocols presented in this paper using smaller neighborhoods.

In [4], we described a non-adaptive membership protocol for iTrust. In this paper, we present three additional membership protocols for iTrust, namely, the retry, adaptive and combined adaptive membership protocols. We compare these membership protocols with respect to four performance metrics, namely, membership accuracy, match probability, response time and message cost, for various parameter values when the membership churn is high and when the membership is stable.

The novel contributions of this paper are:

- Distributed membership protocols for unstructured peer-to-peer networks, such that each node locally determines its view of the membership without communicating additional messages.

- Approximate membership protocols for unstructured peer-to-peer networks, in which each node maintains its own local view of the membership, which it tries to keep close to the actual membership.
- Distributed local estimators for the size of the membership and the membership churn.
- Non-adaptive, retry, adaptive and combined adaptive membership protocols.

The interested reader can find the source code, the user manuals and additional papers on iTrust at http://itrust.ece.ucsb.edu.

2 Related Work

Mischke and Stiller [17] have characterized peer-to-peer networks for distributed search and retrieval as structured or unstructured. The iTrust system is based on an unstructured network, like Gnutella [10], with random distribution of the metadata and the requests to $2\sqrt{N}$ nodes, like the system of Lv et al. [14].

Prior work on membership has focused on forming an agreed accurate membership, and is typically based on a consensus algorithm. Chandra et al. [2] have shown that it is impossible to achieve an agreed accurate membership in the presence of unreliable processors and unreliable communication. Chockler et al. [3] provide a comprehensive survey of membership protocols and group communication systems, and of their formal specifications. Schiper and Toueg [22] provide an elegant formalization of the membership problem that distinguishes between maintaining and agreeing on a set of members and determining which processes are working and should be members. The iTrust membership protocols do not aim to achieve an agreed accurate membership based on a consensus algorithm. Rather, they allow each member to have its own local view of the membership, and they aim to keep that local view close to the actual membership, with a much lower message cost than consensus-based membership protocols.

Ganesh et al. [9] present a membership service, named SCAMP, for gossip-based protocols that operates in a decentralized and self-configured manner, where no peer has global knowledge of the membership. A node that wishes to join (leave) the membership notifies some nodes in the network to add (remove) it to (from) their views. To prevent a node from becoming isolated, a node periodically tries to discover new nodes if it does not receive any messages for a given period of time. Compared to SCAMP, the iTrust membership protocols place more emphasis on maintaining a node's local view of the membership when the membership churn is high.

Zage et al. [25] present a network-aware and distributed membership protocol that biases neighbor selections towards beneficial nodes, based on multiple system metrics and social network patterns. They demonstrate the effectiveness of their protocol for a network with a high churn rate, through emulation. In the iTrust membership protocols, the nodes do not maintain their views of the membership through biased neighbor selections, which might allow malicious nodes to subvert the membership. Rather, they discover newly joining nodes and detect leaving nodes through the normal course of distributing metadata and requests, which reduces the message cost.

Liu et al. [13] describe a novel age-based membership protocol with a conservative neighbor maintenance scheme under churn, to retain desirable properties such as a low network diameter and a low clustering coefficient. Thus, a bootstrapping node recommends, to a newly joining node, only the nodes that have remained in its view for a long period of time. However, with their protocol, a newly joining node might not discover other nodes very quickly, whereas an older node might have knowledge of a larger number of other nodes. In the iTrust membership protocols, a

boostrapping node sends its entire membership to a newly joining node, regarding all nodes as equals.

Voulgaris et al. [24] present a membership management protocol, named CYCLON, for unstructured peer-to-peer networks, in which each node maintains a small and fixed-size neighbor list. They describe a shuffling protocol for large networks and provide an experimental analysis in which they examine the clustering coefficient and the node degree distribution. The iTrust membership protocols differ from CYCLON in that each node tries to discover as many nodes as possible to include in its local view. In other work [1], we have also investigated the use of neighborhoods and de-clustering (shuffling) for very large iTrust networks.

In BubbleStorm [12, 23], when a node joins the network, it finds an existing connection between two peers and interposes itself between them. When a node leaves the network, it re-connects those two peers before it leaves. If a node crashes, a neighboring peer adds a connection to the other peer when it discovers the crashed node. Thus, BubbleStorm aims to maintain a fixed node degree at all of the nodes in the network. The iTrust membership protocols do not try to maintain a fixed node degree but, rather, allow each node to maintain its own local view of the membership.

PlanetP [7] uses a global index that describes all of the peers and their metadata in a Bloom filter, which it replicates throughout the network using gossiping. iTrust does not use gossiping to distribute the entire membership but, rather, allows each node to maintain its own local view by discovering newly joining nodes and detecting leaving (non-operational) nodes through the normal course of operation of the iTrust messaging protocol.

Richardson and Cox [21] provide an overlay of indices to achieve search and ranking in an unstructured peer-to-peer network. A malicious attack might distort the ranking, by not reporting relevant documents (censorship), by increasing the rank of selected documents, or by not reporting highly ranked documents. The authors discuss the distributed local estimation of system-wide metrics, such as average document length, which are expected to follow a normal distribution. Attempts to distort the metrics introduce skew into the distribution, which they can detect. Their work is somewhat similar to our work [18] on search and ranking in iTrust.

Reiter et al. [20] and Freedman et al. [8] describe anonymous communication layers, with associated membership algorithms. Freenet [6] maintains a semi-structured network with a small world topology for better routing that also provides anonymity. We plan to investigate whether such anonymity schemes are appropriate for iTrust. However, anonymity by itself is not enough to prevent censorship.

Gramoli et al. [11] and Pruteanu et al. [19] use a strategy for estimating churn similar to the strategy we use for iTrust. To refine their churn estimates, they exchange churn estimates between neighbors, which exposes those estimates to malice. In contrast, iTrust does not exchange churn estimates and, thus, is more resistant to malice; even so, the churn estimates of iTrust are quite accurate.

3 iTrust Messaging Protocol

First, we briefly describe the iTrust messaging protocol, because the iTrust membership protocols are dependent on it. Some of the nodes in the membership, referred to as the source nodes, produce information, and make that information available to other nodes. Other nodes in the membership, referred to as the requesting nodes, make requests (queries) and retrieve information from the source nodes.

The steps involved in the iTrust messaging protocol are given below and are illustrated in Figure 1.

1. A source node produces metadata that describes its information, and distributes the metadata, along with the URL of the matching document, to a subset of nodes randomly chosen from its local view of the membership.
2. A requesting node generates a request (query) that contains keywords, and distributes its request to a subset of nodes randomly chosen from its local view of the membership.
3. When a node receives a request containing keywords that match metadata it holds (referred to as an encounter or a match), the node returns, to the requesting node, the URL of the matching document at the source node.
4. The requesting node then uses the URL, provided by the matching node, to retrieve the document from the source node.

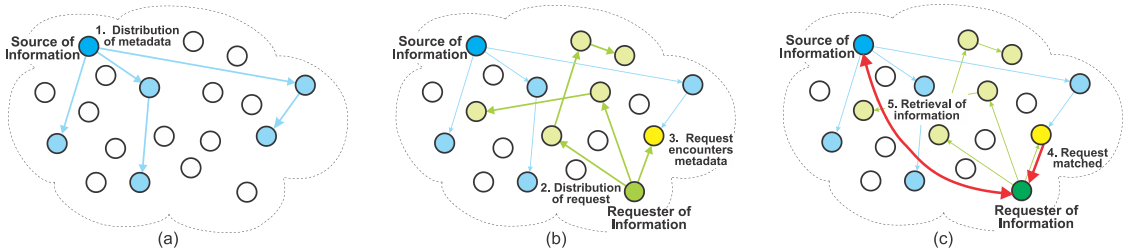


Figure 1. (a) A source node distributes metadata, describing its information, to a subset of nodes randomly chosen from its local view of the membership. (b) A requesting node distributes its request to a subset of nodes randomly chosen from its local view of the membership. (c) One of the nodes matches the metadata and the keywords in the request, and reports the match to the requesting node, which then retrieves the information from the source node

A match between the keywords in a request received by a node and the metadata held by a node might be an exact match or a partial match, or it might correspond to synonyms.

Distribution of the metadata and the requests to relatively few nodes suffices to achieve a high probability of a match. In an iTrust membership with N nodes, distribution of the metadata to $M = \lceil 2\sqrt{N} \rceil$ nodes and distribution of the requests to $R = \lceil 2\sqrt{N} \rceil$ nodes results in a probability of a match that exceeds 0.9817, derived in [15] from the hypergeometric formula for the probability of a match. Distribution of the metadata and the requests to more nodes would result in a higher message overhead, with little improvement in the match probability.

4 iTrust Menbership Protocols

The iTrust messaging protocol, described in Section 3, for metadata and request distribution and for matching and document retrieval depends on a membership, but iTrust does not require an agreed accurate membership, as do some other distributed systems [3]. Rather, iTrust allows each member to have its own local view of the membership, but aims to keep that local view close to the actual membership. We present below the basics of the iTrust membership protocols and, in Sections 6-9, we consider four specific membership protocols, namely the non-adaptive, retry, adaptive and combined adaptive membership protocols. The two adaptive membership protocols utilize the Exponential Weighted Moving Average (EWMA) method. We compare the effectiveness of the four membership protocols when the membership is subject to churn and when the membership is stable.

4.1 Joining the Membership

To join the membership, a node must first obtain the address of a bootstrapping node. To obtain the address of a bootstrapping node, the node uses mechanisms outside the iTrust network, such as conventional Web search, e-mail, Twitter, printed publications, etc.

The steps involved when a node joins the membership are given below, and are illustrated in Figure 2(a).

1. A joining node contacts a bootstrapping node to obtain the bootstrapping node's current view of the membership. A node typically uses a bootstapping node that is known to, and trusted by, the user. Using a malicious bootstrapping node can lead to a seriously distorted membership.
2. The joining node then publishes its joining the membership to a subset of nodes randomly chosen from the view of the membership it obtained from the bootstrapping node.
3. The randomly chosen nodes then add the new node to their local views of the membership.

Another node learns about a new node when it receives a response from a node that is aware of the new node.

4.2 Leaving the Membership

A node may leave the membership either voluntarily, or because it is faulty or disconnected. The steps involved in leaving the membership are simple:

1. To leave the membership, a node simply leaves, without publishing its leaving.

Over time, each node individually discovers the departure of nodes when it sends requests to nodes that do not respond. It is not appropriate to allow a node to publish the departure of another node, because doing so might enable a malicious node to cause the removal of many nodes from the local views of other nodes.

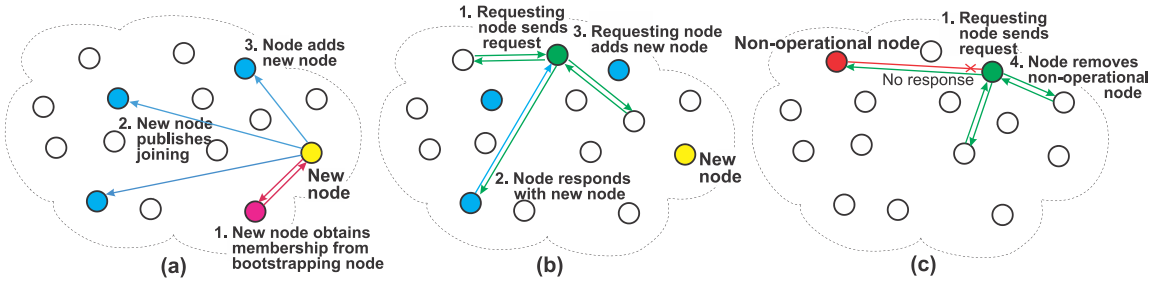


Figure 2. (a) A node joins the membership by first contacting a bootstrapping node to obtain that node's current view of the membership, and then publishing its joining to randomly chosen nodes in that local view. (b) A requesting node distributes a request to nodes randomly chosen from its view of the membership. A node that receives the request returns the identifier(s) of the node(s) that it recently added to its view. A matching node also returns the URL of the document to the requesting node. (c) The requesting node does not receive a response to its request from a node. The requesting node sees a timeout expire or receives an error code from TCP, and then removes that node from its view of the membership.

4.3 Updating the Membership

In the iTrust messaging protocol described in Section 3, each requesting node expects to receive response messages from only the matching nodes. Other nodes that don't have a match are not required to send a response to the requesting node. We have modified that messaging protocol to

enable a requesting node to detect non-operational (leaving) nodes and to discover newly joining nodes, from the responses to its requests.

Now, the requesting node expects each node to which it sent a request to respond with its recently joined member(s), regardless of whether or not that node has a match. Thus, a matching node sends in its response to the requesting node not only the URL of the document at the source node but also the identifier(s) of its recently joined member(s). If it does not have a match, the node responds to the requesting node with the identifier(s) of its recently joined member(s). Thus, the requesting node discovers not only the URLs of the documents, but also newly joined nodes, from the responses to its requests.

If the requesting node doesn't receive a response from a node within a timeout or it receives an error code from TCP, then the non-responding node is considered to have left the membership voluntarily or to be faulty or disconnected, and the requesting node removes that node from its local view of the membership.

The steps involved in updating a requesting node's local view of the membership are given below, and are illustrated in Figure 2(b) and 2(c).

1. A requesting node distributes its request to a subset of nodes randomly chosen from its local view of the membership.
2. A node that receives a request compares the keywords in the request with the metadata it holds. If it finds a match, the node responds to the requesting node with a message that contains the URL of the matching document and also the identifier(s) of its recently joined member(s). A node that doesn't find a match responds to the requesting node with a message that contains the identifier(s) of its recently joined member(s).
3. When the requesting node receives the responses, it adds the new members obtained from the other nodes to its local view of the membership.
4. If the requesting node does not receive a response to its request before a timeout occurs, or if it receives an error code from TCP, then the non-responding node is considered to have left the membership voluntarily or to be faulty or disconnected, and the requesting node removes that node from its local view of the membership.

If the requesting node is also a source node then, after receiving the responses to its request, it distributes its metadata with the URL of the corresponding document to additional nodes, according to the following steps:

1. The requesting node (which is also a source node) calculates the number of nodes to which to distribute its metadata, based on its current view of the membership.
2. Next, the requesting node subtracts the number of nodes to which it previously distributed metadata from the calculated number.
3. Finally, the requesting node distributes its metadata to that many additional nodes, randomly chosen from its current view, but to which it had not sent the metadata previously.

For example, suppose that a requesting node currently has N = 1024 nodes in its local view. It distributes its request to $R = \lceil 2\sqrt{1024} \rceil = 64$ randomly chosen nodes. Suppose further that only 58 nodes reply to the requesting node. From these responses, the requesting node detects that there are 64-58 = 6 non-operational nodes. Suppose that, as a result of receiving the responses from the 58 nodes, the requesting node adds 40 new nodes to its local view. Consequently, the requesting node now has N = 1024-6+40 = 1058 nodes in its view. If the requesting node is also a source node,

then it distributes its metadata to 2√1058 - 2√1024 ~ 65-64 = 1 more node randomly chosen from its new view of the membership.

5 Foundations and Experimental Method

5.1 Environmental Variables

Membership churn refers to nodes joining and leaving the membership, and is represented by the following rates:

- JR: The Joining Rate, the number of nodes that join the membership per time unit. For example, JR = 50 means that 50 nodes join the membership per time unit.
- LR: The Leaving Rate, the number of nodes that leave the membership per time unit. For example, LR = 50 means that 50 nodes leave the membership per time unit.

When the membership has a lot of churn, both *JR* and *LR* are high. When the membership is stable, both *JR* and *LR* are low. These rates are an important consideration for the membership protocols. A node can't control or alter *JR* or *LR*, but it can adjust its requesting rate.

5.2 Parameters for the Membership Protocols

The parameters for the membership protocols are:

- *N*: The number of nodes in a node's local view of the membership.
- *LastJ*: The Last Joined members, the number of recently joined members that a node may report to the requesting node. For example, *LastJ* = 2 allows a node to report its two most recently joined members.
- *Try*: The number of times that a requesting node sends its request message, in an attempt to receive responses from $\lceil 2\sqrt{N} \rceil$ nodes. Because some request messages might be sent to non-operational nodes, a requesting node might need to try several times before it receives responses from $\lceil 2\sqrt{N} \rceil$ nodes.
- *TryMax*: The Maximum Try value, i.e., the maximum number of times that a requesting node is allowed to try to send its request message.
- *RR*: The Requesting Rate, the number of times a node sends a request message to $R = \lceil 2\sqrt{N} \rceil$ nodes per time unit. For example, *RR* = 10 means that a node sends 10 distinct request messages per time unit, each of which is sent to $R = \lceil 2\sqrt{N} \rceil$ nodes.
- *RRMin*: The Minimum Requesting Rate, the minimum rate at which a node is allowed to make requests.
- *RRMax*: The Maximum Requesting Rate, the maximum rate at which a node is allowed to make requests.
- *c*: The weighting factor of the Exponential Weighted Moving Average algorithm used by the adaptive membership protocols.

When a node joins the membership, it obtains the values of LastJ, TryMax, RRMin, RRMax and c. These parameters are tunable for the particular network environment.

5.3 Performance Metrics

The performance metrics for the membership protocols are defined in terms of the following quantities:

- L: The number of leaving nodes that a requesting node hasn't detected.

- J: The number of joining nodes that a requesting node hasn't discovered.
- I: The number of nodes in the intersection of the requesting node's current view of the membership and the actual membership.

The requesting node's current view of the membership consists of I+L nodes, whereas the actual membership consists of I+J nodes. Figure 3 illustrates the quantities I, L and J.

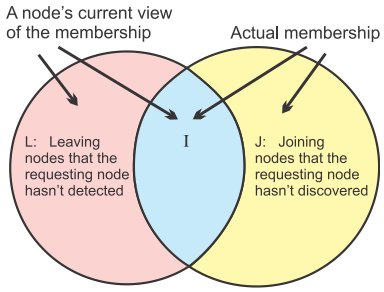


Figure 3. A node's current view of the membership vs. the actual membership

The performance metrics for the membership protocols are:

- *LND*: The Leaves Not Detected, the proportion of leaving (non-operational) nodes in its local view that a requesting node has not detected at a particular time, defined by:

$$LND = \frac{L}{I+L} \quad (1)$$

- *J*ND: The Joins Not Discovered, the proportion of newly joined nodes in the actual membership that a requesting node has not discovered at a particular time, defined bv:

$$JND = \frac{J}{I+J} \quad (2)$$

- *MA*: The Membership Accuracy, the number of nodes in a node's current view that are in the actual membership, divided by that number of nodes plus the number of leaving nodes not detected plus the number of joining nodes not discovered, defined by:

$$MA = \frac{I}{I+L+J} \quad (3)$$

Note that $MA = I/(I+L+J) = 1 - (L+J)/(I+L+J)$, where $L+J$ is the number of leaving nodes that the node hasn't detected plus the number of newly joining nodes that the node hasn't discovered and, therefore, $(L+J)/(I+L+J)$ represents the inaccuracy in the node's local view of the membership.

- *MP*: The Match Probability of one or more responses for a request, averaged over all requesting nodes.
- *RT*: The Response Time for a request, from the time a node starts sending a request to other nodes until it has received all responses, including responses for multiple tries, averaged over all requesting nodes.
- *MC*: The Message Cost per node per time unit, calculated as an average over all nodes over time.

5.4 Measured Values

The membership protocols for iTrust use the following measured values:

- *Left*: The number of nodes that a requesting node has detected to have left the membership since its last request.
- *Joined*: The number of nodes that a requesting node has discovered to have joined the membership since its last request.

- *NumNodes*: The number of nodes to which the requesting node sent its request.

Using these measured values, for each of the two adaptive membership protocols, a node calculates a *Churn Estimator* for each request when it finishes receiving the responses to that request, defined as follows:

- CE: The Churn Estimator, an estimate of the leaves and joins (churn) obtained by random sampling, given by:

$$CE = \frac{Left + Joined}{NumNodes} \tag{4}$$

The Churn Estimator is used by the adaptive membership protocols to adapt the Requesting Rate *RR*. The values of the Churn Estimator are determined using the Exponential Weighted Moving Average algorithm, described below.

5.5 Exponential Weighted Moving Average Algorithm

The adaptive membership protocols for iTrust uses the Exponential Weighted Moving Average (EWMA) algorithm to process a sequence of estimated values of the Churn Estimator *CE*, to smooth the estimated values and to reduce the noise inherent in the individual samples.

A requesting node issues requests (queries), collects responses, detects non-operational nodes, and discovers newly joined nodes. It then computes the estimated value *CE*, using the EWMA algorithm defined by:

$$\begin{aligned} s_1 &= v_1 \\ s_t &= c \times v_t + (1\text{-}c) \times s_{t-1} \quad \text{if } t > 1 \end{aligned} \tag{5}$$

where v_t is the measured value at time t and s_t is the smoothed value at time t. The constant c is a smoothing factor, $0 < c < 1$. Larger values of c place more emphasis on the most recently measured values and a faster response to changes. Smaller values of c provide more smoothing and less vulnerability to random fluctuations and noise.

The pseudocode for the EWMA algorithm is given in Figure 4.

```
EWMA(v, c, s)
1   if (t = 0) then s ← v
2   else s ← c × v + ( 1 – c ) × s
3   return s
```

Figure 4. Pseudocode for the EWMA algorithm

In our experiments, we used $c = 0.7$. However, iTrust offers the user the option to choose a value of c for the user's particular network environment. Different users, who operate in different network environments with different objectives, may choose different values of c, which iTrust allows.

5.6 Experimental Method

To evaluate and compare the four membership protocols for iTrust, described in Sections 6-9, we performed experiments using an emulation of iTrust. In the emulation, we can control the Leaving Rate *LR* and the Joining Rate *JR*, which a real-world deployment would not allow us to do. Moreover, in the emulation, we can compare a node's current view of the membership with the actual membership.

The emulation of iTrust is based on our HTTP implementation of iTrust using the Apache Web server, running on Debian Linux 6.0 on an Intel Quad Core 3.4 GHz processor with 4 GB memory and 1 TB

hard drive. In the emulation, we have multiple virtual hosts installed on a single Apache Web server, where each virtual host represents a node in the iTrust network. Each node has a separate SQLite database that resides on the Apache Web server, where it stores queries and resource information.

Before we start the emulation program, we set the value of *N*, the number of nodes in the initial membership. The program clears the node's resources and databases, and then adds all of the nodes to each node's view of the membership, so that each node has the complete initial membership. At each time step, nodes might join the membership, leave the membership, and make requests. Different nodes might have different views of the membership, and different nodes might make requests at different rates.

For each source node, iTrust creates metadata for a document that the node wishes to share, and distributes the metadata to *M* randomly chosen nodes in the membership set. Then, iTrust distributes requests to *R* randomly chosen nodes in the membership set. Finally, the program compares each node's view of the membership against the actual membership. The program computes the four performance metrics at each time step.

6 Non-Adaptive Membership Protocol

The Non-Adaptive Membership Protocol implements the membership protocol, described in Section 4, which involves requesting nodes updating their local views of the membership, as other nodes join and leave the membership.

The pseudocode for the Non-Adaptive Membership Protocol is given in Figure 5. The inputs for the Non-Adaptive Membership Protocol are *N* and *RR*, where *N* is the number of nodes in the node's local view and *RR* is the node's requesting rate.

The Non-Adaptive Membership Protocol comprises an infinite loop, at the beginning of which *nextRequestTime* is set to the current *time* plus *timeunit*/*RR*, which is the time when the node sends its next request. As time passes, *time* is automatically incremented (not shown in the pseudocode); *timeunit* is the length of the time unit. The protocol waits (line 3) until the current *time* reaches the *nextRequestTime*.

```
NonAdaptive(N, RR)
1  while true do
2      nextRequestTime ← time + (timeunit/RR)
3      wait until (time = nextRequestTime)
4      R ← ⌈2×√N⌉
5      responses ← makeRequests(view, R)
6      responded ← 0
7      for ( j ← 0 to R) do
8          if (responses[j].noResponse) then
9              removeNode(view, responses[j].node)
10             N ← N – 1
11         else
12             responded ← responded + 1
13             for (k ← 0 to LastJ) do
14                 if (responses[j].recent[k]) then
15                     isNew ← addNode(view, responses[j].recentNode[k])
16                     if (isNew) then
17                         N ← N + 1
```

Figure 5: Non-Adaptive Membership Protocol

The protocol then sets the number *R* of nodes to which the node sends its request message to $\lceil 2\sqrt{N} \rceil$, where *N* is the number of nodes in the node's current local view. The node then sends its request to *R* nodes, and waits for responses from those nodes (line 5).

Next, the protocol iterates through the *responses* array (line 7). The protocol checks whether the node received a response from node *j*. If not, it removes the non-responsive node from the node's local view and then decrements the number *N* of nodes in that view. Otherwise the protocol increments *responded*.

The protocol then iterates (line 13) through the *responses*[*j*].*recent* array and the *responses*[*j*].*recentNode* array, which provides the identifiers of up to the *LastJ* recently joined nodes. It checks whether *responses*[*j*].*recent*[*k*] is true. If so, it invokes *addNode*() to add the recent node *responses*[*j*].*recentNode*[*k*] to its local view. The procedure *addNode*() returns a boolean *isNew* to indicate whether or not the recent node is already present in the node's local view. If the recent node is indeed new, the protocol increments the number *N* of nodes in the node's local view.

Control then returns to continue the iteration through the *recentNode* array (line 13) or the *responses* array (line 7). When the protocol finishes iterating through the *responses* array, it goes back to the beginning, and repeats these steps indefinitely.

6.1 Investigation of the Non-Adaptive Membership Protocol

We investigate the Non-Adaptive Membership Protocol, in particular, the number *LastJ* of newly joined nodes that a responding node may report to a requesting node, and its effect on *LND* (the proportion of leaving nodes that the requesting node has not detected) and *JND* (the proportion of newly joined nodes that the requesting node has not discovered).

In the Non-Adaptive Membership Protocol, a requesting node distributes its request to $R = \lceil 2\sqrt{N} \rceil$ nodes chosen at random from its local view. Initially, we required those *R* nodes to return their entire views to the requesting node, and the requesting node to update its local view accordingly. The problem with that approach is that the requesting node adds to its view non-operational nodes obtained from other nodes that haven't yet detected that those nodes are non-operational. Thus, the requesting node adds to its local view non-operational nodes, including nodes that it recently removed.

Several possible solutions to this problem exist. One solution is that, once a requesting node has obtained the views of the other nodes, it sends a "verify" message to confirm that those nodes are indeed operational. Such a solution consumes too much network bandwidth. An alternative, less costly solution is to require the $\lceil 2\sqrt{N} \rceil$ nodes to return, to the requesting node, their "most recently joined members," rather than their entire views. We adopt the latter solution and investigate how *LastJ* affects *LND* and *JND*.

Consider a scenario where *N* = 1024 nodes with a high leaving rate (*LR* = 300), a high joining rate (*JR* = 300) and a low requesting rate (*RR* = 10). Figure 6 shows the graphs for *LND* and *JND* over time for this scenario. Increasing *LastJ* from *LastJ* = 1 in the left graph to *LastJ* = 2 in the middle graph results in a decrease in *JND* but an increase in *LND*. Increasing *LastJ* from *LastJ* = 2 in the middle graph to *LastJ* = 3 in the right graph results in a slight decrease in *JND* and little change in *LND*.

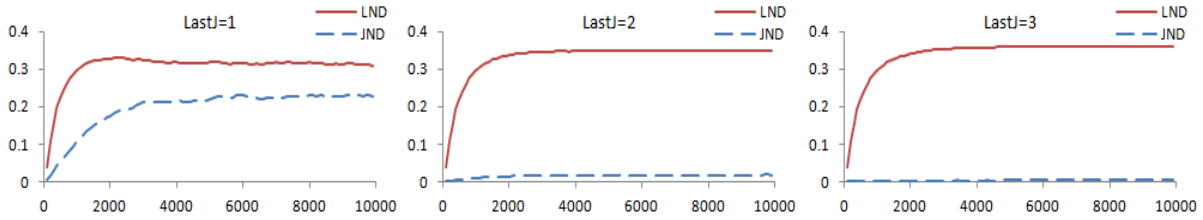


Figure 6: Graphs for the Non-Adaptive Membership Protocol, showing *LND* and *JND* over time for *LastJ* = 1, *LastJ* = 2 and *LastJ* = 3, where *N* = 1024 initially, *LR* = 300, *JR* = 300 and *RR* = 10

Thus, increasing *LastJ* definitely helps the requesting node to discover more joining nodes as it issues more requests. However, increasing *LastJ* also causes the requesting node to add back into its local view too many non-operational (leaving) nodes. Setting *LastJ* = 2 or *LastJ* = 3 increases *LND* with worse results than setting *LastJ* = 1. We conclude that *LastJ* = 1 is an appropriate value to use in our further experiments.

7 Retry R Membership Protocol

As we have seen, increasing *LastJ* does not help to detect leaving (non-operational) nodes. Thus, we investigate other methods to reduce the proportion *LND* of leaving nodes that a requesting node has not detected.

When a node distributes a request message to $\lceil 2\sqrt{N} \rceil$ nodes, it might not receive $\lceil 2\sqrt{N} \rceil$ responses for its request, because non-operational (leaving) nodes do not respond. Thus, we investigate a retry method that allows the requesting node to distribute its request to more than $\lceil 2\sqrt{N} \rceil$ nodes until it receives $\lceil 2\sqrt{N} \rceil$ responses. We call this protocol the Retry *R* Membership Protocol.

The pseudocode for the Retry *R* Membership Protocol is given in Figure 7. The inputs for this protocol are *N*, *TryMax* and *RR*, where *N* is the number of nodes in the node's current local view, *TryMax* is the number of times a node is allowed to try to obtain $\lceil 2\sqrt{N} \rceil$ responses and *RR* is the node's requesting rate.

The Retry *R* Membership Protocol comprises an infinite loop, at the beginning of which *nextRequestTime* is set to the current *time* plus the time *timeunit*/*RR* until the next request. The protocol waits (line 3) until the current *time* reaches the *nextRequestTime*.

The protocol then sets the number *R* of nodes to which the node sends its request message to $\lceil 2\sqrt{N} \rceil$. It sets the number *resRec* of responses received to 0 and the number *Try* of tries to 1, and then starts the while loop (line 7). The node sends its request message to *R* - *resRec* nodes, and waits for responses from those nodes (line 8).

Next, the protocol iterates through the *responses* array (line 10). It checks whether the node received a response from node *j*. If not, it removes the non-responsive node from the node's local view and then decrements the number *N* of nodes in that view. Otherwise, it increments *responded*.

The protocol then checks whether *j*'s response contains a recent node. If so, it invokes *addNode*() to add the recent node to the node's local view. The protocol then checks whether the recent node is indeed new and, if so, it increments the number *N* of nodes in the node's view. Control then returns to continue the iteration through the *responses* array (line 10).

```
RetryR(N, TryMax, RR)
1  while true do
2      nextRequestTime ← time + (timeunit / RR)
3      wait until (time = nextRequestTime)
4      R ← ⌈2×√N⌉
5      resRec ← 0
6      Try ← 1
7      while ((Try ≤ TryMax) and (resRec < R)) do
8          responses ← makeRequests(view, R – resRec)
9          responded ← 0
10         for (j ← 0 to (R – resRec)) do
11             if (responses[j].noResponse) then
12                 removeNode(view, responses[j].node)
13                 N ← N – 1
14             else
15                 responded ← responded + 1
16                 if (responses[j].recent) then
17                     isNew ← addNode(view,responses[j].recentNode)
18                     if (isNew) then
19                         N ← N + 1
20         resRec ← resRec + responded
21         Try ← Try + 1
```

Figure 7: Retry *R* Membership Protocol

After it has finished iterating through the *responses* array, the protocol increases *resRec* by *responded*, the number of responses in the *responses* array and then increments *Try*. Control then returns to the while loop (line 7) to determine whether *Try* is less than or equal to *TryMax* and *resRec* is less than *R*. If both of these conditions are satisfied, the protocol goes through the while loop again. Otherwise, the protocol goes back to the beginning, and repeats these steps indefinitely.

7.1 Investigation of the Retry R Membership Protocol

We investigate the Retry *R* Membership Protocol, in particular, the number *Try* of times that a requesting node tries to send its request message, in order to receive $\lceil 2\sqrt{N} \rceil$ responses. For example, *Try* = 1 means that a requesting node sends its request message to $\lceil 2\sqrt{N} \rceil$ nodes regardless of the number of responses that it receives to its request. *Try* = 2 means that a requesting node tries a second time and sends its request to *Left* nodes, where *Left* nodes didn't respond on the first Try and, similarly, for *Try* = 3. *Try* = ∞ means that a requesting node sends its request repeatedly until it receives responses from $\lceil 2\sqrt{N} \rceil$ nodes for that particular request.

Table 1 shows the membership accuracy, match probability, response time and message cost for *Try* = 1, 2, 3, ∞, where *N* = 1024 initially, *LastJ* = 1, *LR* = 300, *JR* = 300 and *RR* = 10 for the Retry *R* Membership Protocol.

Table 1: Retry *R* Membership Protocol with *Try* = 1, 2, 3, ∞

Try	1	2	3	∞
Membership Accuracy	0.5966	0.6821	0.6986	0.7048
Match Probability	0.9345	0.9817	0.9865	0.9864
Response Time	6.0	11.9274	17.6573	24.1682
Message Cost	3.8552	5.1538	5.4527	5.5598

In Table 1, we see that, as *Try* is increased from *Try* = 1 to *Try* = 2, both the membership accuracy and the match probability are greatly increased but, when *Try* is further increased to *Try* = 3, there is not much increase in either the membership accuracy or the match probability. We also see that both the response time and the message cost increase as *Try* is increased. To obtain a substantial increase in the membership accuracy and the match probability with a reasonable increase in the response time and the message cost, we use *Try* = 2 in our further experiments.

8 Adaptive RR Membership Protocol

The next membership protocol we consider uses the Churn Estimator *CE* to control the Requesting Rate *RR*. The Churn Estimator *CE* provides an estimate of the churn (leaves and joins) in the network; it is initialized to 0, and the *CE* values are averaged using the Exponential Weighted Moving Average (EWMA) algorithm as time progresses. We call this protocol the Adaptive *RR* Membership Protocol.

The pseudocode for the Adaptive *RR* Membership Protocol is given in Figure 8. The inputs for the Adaptive *RR* Membership Protocol are *N*, *RR*, *RRMin*, *RRMax* and *c*. Here *N* is the number of nodes in the node's current local view, and *RR* is the node's initial requesting rate.

```
AdaptiveRR(N, RR, RRMin, RRMax, c)
1  CE ← 0
2  while true do
3      nextRequestTime ← time + (timeunit/RR)
4      wait until (time = nextRequestTime)
5      Left ← 0
6      Joined ← 0
7      R ← ⌈2×√N⌉
8      responses ← makeRequests(view, R)
9      for (j ← 0 to R) do
10         if (responses[j].noResponse) then
11             removeNode(view, responses[j].node)
12             N ← N - 1
13             Left ← Left + 1
14         else
15             if (responses[j].recent) then
16                 isNew ← addNode(view, responses[j].recentNode)
17                 if (isNew) then
18                     N ← N + 1
19                     Joined ← Joined + 1
20     currentCE ← (Left + Joined) / R
21     CE ← EWMA(CE, currentCE, c)
22     if CE > RRMin / RRMax then
23         RR ← RRMax × CE
24     else
25         RR ← RRMin
```

Figure 8: Adaptive *RR* Membership Protocol

The Adaptive *RR* Membership Protocol comprises an infinite loop, at the beginning of which *nextRequestTime* is set to the current *time* plus *timeunit*/*RR*, until the next request. The protocol waits (line 4) until the current *time* reaches the *nextRequestTime*.

The protocol initializes the variables *Left* and *Joined* (which count the changes in the node's view) to 0, and sets the number *R* of nodes to which the node sends its request message to $\lceil 2\sqrt{N} \rceil$. The node then sends its request to *R* nodes, and waits for their responses (line 8).

Next, the protocol iterates through the *responses* array (line 9). It checks whether the node received a response from node *j*. If not, it removes the non-responsive node from the node's local view and then decrements the number *N* of nodes in that view and increments *Left*, the number of nodes that have left. Otherwise, the protocol checks whether *j*'s response contains a recent node. If so, the protocol invokes *addNode*() to add the recent node to the node's local view. The protocol then checks whether the recent node is indeed new and, if so, it increments the number *N* of nodes in the

node's local view and *Joined*, the number of nodes that have recently joined. Control then returns to continue the iteration through the *responses* array (line 9).

After processing the *responses* array, the protocol calculates *currentCE* (line 20) and then applies the EWMA algorithm to obtain the smoothed value of the Churn Estimator *CE* (line 21). The protocol then calculates the value of the Requesting Rate *RR* for the next time unit, corresponding to the smoothed value of the Churn Estimator *CE*. It then goes back to the beginning of the loop and repeats these steps indefinitely.

8.1 Investigation of the Retry R Membership Protocol

Figure 9 shows the graphs for the Leaves Not Detected (*LND*) and the Joins Not Discovered (*JND*) over time for the Non-Adaptive (left graph), the Retry *R* (middle graph) and the Adaptive *RR* (right graph) Membership Protocols. Here *N* = 1024 initially, *LastJ* = 1, *LR* = 300, *JR* = 300 and *c* = 0.7. For the Non-Adaptive and the Retry *R* Membership Protocols, *RR* = 10. For the Adaptive *RR* Membership Protocol, *RRMin* = 1, *RRMax* = 100 and *RR* = 10 initially.

In the middle graph of Figure 9, we see that, for the Retry *R* Membership Protocol, *LND* decreases to about 0.26 and *JND* decreases to about 0.12. In the right graph, we see that for the Adaptive *RR* Membership Protocol, *LND* greatly decreases to about 0.14 and *JND* decreases to almost 0. Thus, from these graphs, we see that increasing *RR* is more effective than increasing the number of tries, in decreasing both *LND* and *JND*.

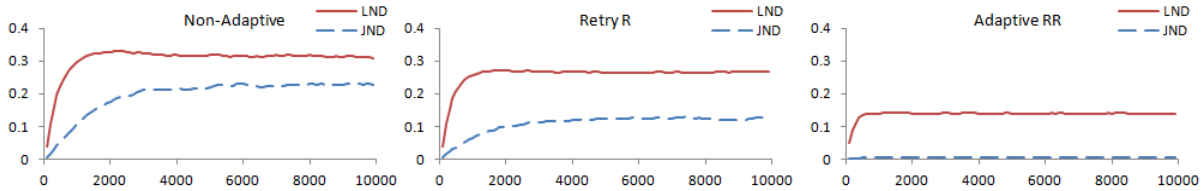


Figure 9: Graphs for the Non-Adaptive, the Retry *R* and the Adaptive *RR* Membership Protocols, showing Leaves Not Detected (*LND*) and Joins Not Discovered (*JND*), over time, where *N* = 1024 initially, *LastJ* = 1, *LR* = 300 and *JR* = 300

Table 2 presents the membership accuracy, match probability, response time and message cost for the Non-Adaptive, the Retry *R* and the Adaptive *RR* Membership Protocols.

In Table 2, we see that the Adaptive *RR* Membership Protocol has the highest membership accuracy, whereas the Non-Adaptive and the Retry *R* Membership Protocols have a much lower membership accuracy. We also

Table 2: Non-Adaptive vs. Retry *R* with *Try* = 2 vs. Adaptive *RR* with *Try* = 1 and *RRMax* = 100

	Non-Adaptive	Retry *R*	Adaptive
Membership Accuracy	0.5966	0.6821	0.8581
Match Probability	0.9345	0.9817	0.9728
Response Time	6.0	11.9274	6.0
Message Cost	3.8552	5.1538	12.5690

notice that the Retry *R* Membership Protocol achieves the highest match probability, whereas the Non-Adaptive Membership Protocol has the lowest match probability. Both the Non-Adaptive and the Adaptive *RR* Membership Protocols have a low response time, whereas the Retry *R* Membership Protocol has a much higher response time. The Adaptive *RR* Membership Protocol has a message cost that is about 3 times that of the Non-Adaptive Membership Protocol and about 2.5 times that of the Retry *R* Membership Protocol.

Thus, the Adaptive *RR* Membership Protocol has the highest membership accuracy, which is achieved by increasing the Requesting Rate *RR*, resulting in a high message cost. In contrast, the Retry *R* Membership Protocol has the highest match probability with a lower message cost but it also has a lower membership accuracy and a higher response time.

Therefore, we continue our investigations into an adaptive membership protocol that is intermediate between the Retry *R* Membership Protocol and the Adaptive *RR* Membership Protocol, in order to reduce the Requesting Rate *RR* and thus the message cost.

9 Combined Adaptive Membership Protocol

Now, we consider a membership protocol that not only adapts a node's Requesting Rate *RR* based on the Churn Estimator *CE* but also tries a second time (*Try* = 2) to obtain $\lceil 2\sqrt{N} \rceil$ responses to a node's request. We call this protocol the Combined Adaptive Membership Protocol.

The pseudocode for the Combined Adaptive Membership Protocol is given in Figure 9. The inputs for the Combined Adaptive Membership Protocol are *N*, *RR*, *RRMin*, *RRMax* and *c*. Again, *N* is the number of nodes in the node's current view, and *RR* is the node's current requesting rate.

The Combined Adaptive Membership Protocol comprises an infinite loop, in which *nextRequestTime* is first set to the current *time* plus *timeunit*/*RR*, until the next request. The protocol waits (line 4) until the current *time* reaches the *nextRequestTime*.

The protocol initializes the variables *Left* and *Joined* to 0, and sets the number *R* of nodes to which the node sends its request message to $\lceil 2\sqrt{N} \rceil$. It also initializes the number *resRec* of nodes from which it received responses to 0, and initializes the variable *Try* to 1.

The loop commencing at line 10 is potentially executed twice, but only once if the number of responses received (*resRec*) in the first try is equal to the number *R* of nodes to which the node sent its request message. Within the loop, the node sends its request message to *R* - *resRec* nodes, and waits for responses from those nodes (line 11).

The protocol then iterates through the *responses* array (line 13). It checks whether the node received a response from node *j*. If not, the protocol removes the non-responsive node from the node's view and then decrements the number *N* of nodes in its view and increments *Left*, the number of nodes that have left. Otherwise, the protocol increments *responded*. Then, it checks whether *j*'s response contains a recent node. If so, the protocol invokes *addNode*() to add the recent node to the node's view. The protocol then checks whether the recent node is indeed new and, if so, it increments the number *N* of nodes in the node's view and *Joined*, the number of newly joined nodes in that view. The protocol increases *resRec* by *responded*, the number of nodes that responded in this try, and then increments *Try*. Control then returns to the while loop (line 10) to determine whether *Try* is less than or equal to 2 and *resRec* is less than *R*. If both of these conditions are satisfied, the protocol goes through the while loop again.

After the protocol has completed the while loop, it calculates *currentCE* (line 27) using the values of *Left*, *Joined* and *resRec* it obtained in the loop, and then invokes the EWMA algorithm to calculate the smoothed value of the Churn Estimator *CE* (line 28). The protocol then calculates the value of the Requesting Rate *RR* for the next time unit, corresponding to the smoothed value of the Churn Estimator *CE*. The protocol then goes back to the beginning of the loop and repeats these steps indefinitely.

```
CombinedAdaptive(N, RR, RRMin, RRMax, c)
1  CE ← 0
2  while true do
3      nextRequestTime ← time + (timeunit/RR)
4      wait until (time = nextRequestTime)
5      Left ← 0
6      Joined ← 0
7      R ← ⌈2×√N⌉
8      resRec ← 0
9      Try ← 1
10     while ((Try ≤ 2) and (resRec < R)) do
11         responses ← makeRequests(view, R-resRec)
12         responded ← 0
13         for (j ← 0 to (R – resRec)) do
14             if (responses[j].noResponse) then
15                 removeNode(view, responses[j].node)
16                 N ← N – 1
17                 Left ← Left + 1
18             else
19                 responded ← responded + 1
20                 if (responses[j].recent) then
21                     isNew ← addNode(view,responses[j].recentNode)
22                     if (isNew) then
23                         N ← N + 1
24                         Joined ← Joined + 1
25         resRec ← resRec + Responded
26         Try ← Try + 1
27     currentCE ← (Left + Joined) / (R + R – resRec)
28     CE ← EWMA(CE, currentCE, c)
29     if CE > RRMin / RRMax then
30         RR ← RRMax × CE
31     else
32         RR ← RRMin
```

Figure 10: Pseudocode for the Combined Adaptive Membership Protocol

Note that, with *Try* = 2, a requesting node sends its request to more nodes to try to obtain $\lceil 2\sqrt{N} \rceil$ responses to its request. Thus, the Combined Adaptive Membership Protocol does not need to increase the Requesting Rate *RR* as much as does the Adaptive *RR* Membership Protocol. Consequently, the Combined Adaptive Membership Protocol realizes some savings in the message cost, compared to the Adaptive *RR* Membership Protocol.

9.1 Investigation of the Combined Adaptive Membership Protocol

We investigate the Combined Adaptive Membership Protocol, in particular the values of the Maximum Requesting Rate *RRMax*.

Table 3 shows the values of the membership accuracy, match probability, response time and message cost for the Combined Adaptive Membership Protocol for *RRMax* = 100, 50 and 30.

Table 3: Combined Adaptive Membership Protocol with *RRMax* = 100, 50 and 30

RRMax	100	50	30
Membership Accuracy	0.8663	0.8198	0.7579
Match Probability	0.9843	0.9836	0.9822
Response Time	11.9874	11.9883	11.9885
Message Cost	13.4939	9.3156	6.9104

We see that, as *RRMax* is decreased from *RRMax* = 100 to *RRMax* = 50, the membership accuracy decreases and the match probability slightly decreases. When *RRMax* is further decreased to *RRMax* = 30, again the membership accuracy decreases and the match probability slightly decreases, but still remains quite good.

We also see that, as *RRMax* is decreased, the response time remains the same and the message cost decreases substantially. The message cost for *RRMax* = 100 is nearly twice that for *RRMax* = 30. To keep the message cost lower while obtaining good membership accuracy, we chose *RRMax* = 50 for our further experiments.

9.2 Comparison of the Retry R, the Adaptive RR and the Combined Adaptive Membership Protocols

Table 4 shows the membership accuracy, match probability, response time and message cost for the Retry *R*, the Adaptive *RR* and the Combined Adaptive Membership Protocols. Here, *N* = 1024 initially, *LastJ* = 1, *LR* = 300, *JR* = 300 and *c* = 0.7. For the Retry *R* Membership Protocol, *Try* = 2 and *RR* = 10. For the Adaptive *RR* Membership Protocol, *Try* = 1, *RRMin* = 1, *RRMax* = 100 and *RR* = 10 initially. For the Combined Adaptive Membership Protocol, *Try* = 2, *RRMin* = 1, *RRMax* = 50 and *RR* = 10 initially.

Table 4: Retry *R* with *Try* = 2 vs. Adaptive *RR* with *Try* = 1 and *RRMax* = 100 vs. Combined Adaptive with *Try* = 2 and *RRMax* = 50

	Retry *R*	Adaptive *RR*	Combined Adaptive
Membership Accuracy	0.6821	0.8581	0.8198
Match Probability	0.9817	0.9728	0.9836
Response Time	11.9274	6.0	11.9883
Message Cost	5.1538	12.5690	9.3156

From Table 4, we see that the membership accuracy of the Combined Adaptive Membership Protocol is 0.8198, which is much better than that of the Retry *R* Membership Protocol, but worse than that of the Adaptive *RR* Membership Protocol. We also see that the match probability of the Combined Adaptive Membership Protocol is 0.9836, which is good and better than that of the Adaptive *RR* Membership Protocol and slightly better than that of the Retry *R* Membership Protocol. We see further that the response time of the Combined Adaptive Membership Protocol is about the same as that of the Retry *R* Membership Protocol, which is about double that of the Adaptive *RR* Membership Protocol. We also see that the message cost of the Combined Adaptive Membership Protocol lies between that of the Retry *R* Membership Protocol and that of the Adaptive *RR* Membership Protocol, and is about three-fourths that of the Adaptive *RR* Membership Protocol.

In summary, the Combined Adaptive Membership Protocol balances the message cost against the membership accuracy. The message cost of the Combined Adaptive Membership Protocol is less than that of the Adaptive *RR* Membership Protocol and also the membership accuracy of the Combined Adaptive Membership Protocol is greater than that of the Retry *R* Membership Protocol.

10 Extended Scenario

Now we investigate the effectiveness of the Combined Adaptive Membership Protocol to see how well it handles various combinations of low and high Leaving Rate *LR* and low and high Joining Rate *JR*. In particular, we consider an extended scenario that comprises the following five scenarios:

- Scenario 1: *LR* = 10, *JR* = 10 for time 0 to 3000
- Scenario 2: *LR* = 300, *JR* = 300 for time 3000 to 6000
- Scenario 3: *LR* = 0, *JR* = 300 for time 6000 to 9000

- Scenario 4: *LR* = 300, *JR* = 0 for time 9000 to 12000
- Scenario 5: *LR* = 0, *JR* = 0 for time 12000 to 15000.

For all five scenarios, we set *LastJ* = 1 and *c* = 0.7. Initially, there are *N* = 1024 nodes in the membership, and each node's view is the entire membership. As time progresses, each member changes its local view. The number *M* of nodes to which the metadata are distributed and the number *R* of nodes to which the requests are distributed are both set to $\lceil 2\sqrt{N} \rceil$, where *N* is the number of nodes in the node's current view at a given time step.

We compare the effectiveness of the Combined Adaptive Membership Protocol and the Non-Adaptive Membership Protocol by considering the extended scenario that comprises these five scenarios.

10.1 Non-Adaptive Membership Protocol

Figure 11 shows the graphs of the Leaves Not Detected *LND*, Joins Not Discovered *JND*, Membership Accuracy *MA* and Match Probability *MP* for the Non-Adaptive Membership Protocol. Here, *LastJ* = 1 and the Requesting Rate *RR* = 10 for all five scenarios.

In the first scenario of Figure 11, the Leaving Rate *LR*, Joining Rate *JR* and Requesting Rate *RR* are low and the same (*LR* = *JR* = *RR* = 10). The values of *LND* and *JND* remain low, because a node detects non-operational (leaving) nodes and discovers newly joining nodes within a short time interval. The Membership Accuracy *MA* remains high at about 0.9873 throughout the first scenario. The Match Probability *MP* is generally higher than the value 0.9817 obtained from the hypergeometric formula [15].

In the second scenario, the values of the Leaving Rate *LR* and the Joining Rate *JR* are much higher than the value of the Requesting Rate *RR* (*LR* = *JR* = 300 and *RR* = 10). The values of *LND* and *JND* increase, because a node can't detect enough non-operational (leaving) nodes and can't discover enough newly joined nodes within a short time interval. The Membership Accuracy *MA* dramatically decreases to about 0.5852. Moreover, the Match Probability *MP* is quite variable, decreasing to about 0.85 and then increasing to about 0.9350.

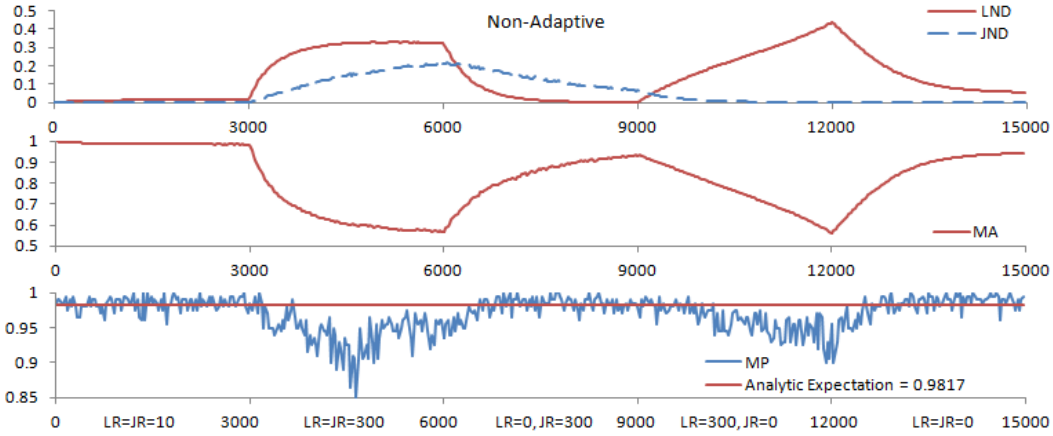


Figure 11: Graphs of *LND*, *JND*, *MA* and *MP* for the Non-Adaptive Membership Protocol where *LastJ* = 1 and *c* = 0.7

In the third scenario, the value of the Leaving Rate *LR* is low, the value of the Joining Rate *JR* is high, and the value of the Requesting Rate *RR* is low (*LR* = 0, *JR* = 300 and *RR* = 10). The values of *LND* decrease because *LR* drops to *LR* = 0. The values of *JND* remain high because *JR* remains high. The Membership Accuracy *MA* is higher than that in the second scenario (because *LR* = 0), and slowly

increases to about 0.9331 at the end of the third scenario. Similarly, the Match Probability *MP* slowly increases from about 0.9350 to about 0.9850.

In the fourth scenario, the value of the Leaving Rate *LR* is high, the value of the Joining Rate *JR* is low, and the value of the Requesting Rate *RR* is low (*LR* = 300, *JR* = 0 and *RR* = 10). The values of *LND* increase to about 0.4384. In addition, the Membership Accuracy *MA* steadily decreases to about 0.5610. The reason is that most of the nodes haven't yet discovered all of the newly joined nodes from the third scenario, but now more nodes are leaving the membership. The Match Probability *MP* fluctuates considerably, decreasing to about 0.9.

Lastly, in the fifth scenario, the values of both the Leaving Rate *LR* and the Joining Rate *JR* are low and the value of the Requesting Rate *RR* is also low (*LR* = *JR* = 0 and *RR* = 10). Thus, the Membership Accuracy *MA* slowly increases to about 0.9426. In addition, the Match Probability *MP* increases and remains high, hovering around the analytic expectation 0.9817.

10.2 Combined Adaptive Membership Protocol

Figure 12 shows the Leaves Not Detected *LND*, Joins Not Discovered *JND*, Requesting Rate *RR*, Membership Accuracy *MA* and Match Probability *MP* for the Combined Adaptive Membership Protocol. Here, *LastJ* = 1, *c* = 0.7, *Try* = 2, *RRMax* = 50 and *RR* = 10 initially.

In the first scenario, the values of the Leaving Rate *LR* and the Joining Rate *JR* are low (*LR* = *JR* = 10). Thus, the values of both *LND* and *JND* are low, because there are not many non-operational (leaving) nodes or newly joined nodes. The value of the Requesting Rate *RR* quickly decreases to 3.6764, in order to reduce the message cost. The Membership Accuracy *MA* remains high throughout the first scenario, and the Match Probability *MP* hovers around the analytic expectation 0.9817.

In the second scenario, the values of the Leaving Rate *LR* and the Joining Rate *JR* are high (*LR* = *JR* = 300), much higher than the values of the Requesting Rate *RR* (*RRMax* = 50). The values of *JND* and *LND* shown in

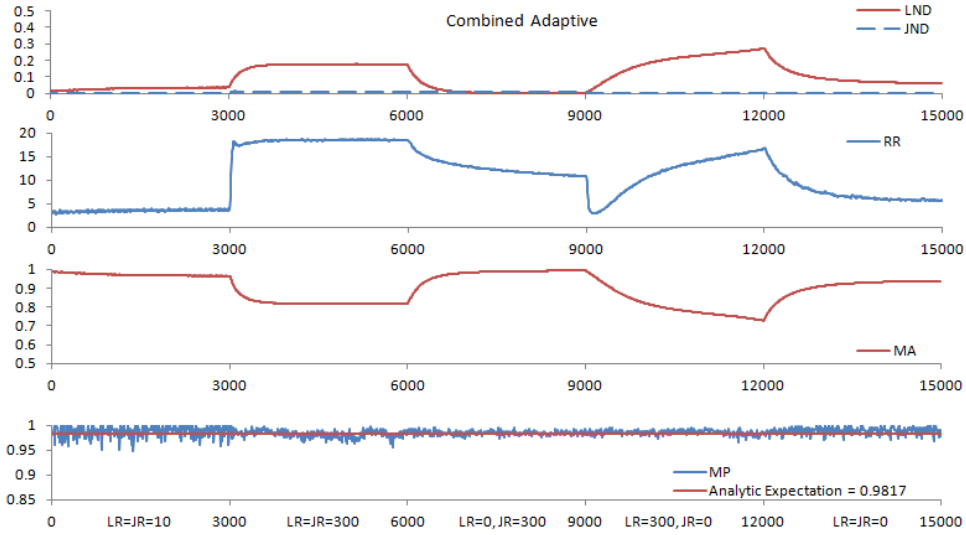


Figure 12: Graphs of *LND*, *JND*, *RR*, *MA* and *MP* for the Combined Adaptive Membership Protocol, where *RR* = 10, *LastJ* = 1, *c* = 0.7, *RRMax* = 50 and *Try* = 2

Figure 12 are much less than the corresponding values for the Non-Adaptive Membership Protocol shown in Figure 11. For the Combined Adaptive Membership Protocol, the value of the Requesting Rate *RR* is increased to about 18.5161, which results in a Membership Accuracy *MA* of about 0.8251 compared to about 0.5852 for the Non-Adaptive Membership Protocol. Lastly, the Match

Probability *MP* remains high throughout the second scenario, hovering around the analytic expectation 0.9817, whereas for the Non-Adaptive Membership Protocol it decreases to about 0.85.

In the third scenario, the value of the Leaving Rate *LR* is low and the value of the Joining Rate *JR* is high (*LR* = 0, *JR* = 300). Because *LR* is low, the values of *LND* remain close to 0. The values of *JND* also remain close to 0. The Combined Adaptive Membership Protocol adjusts the value of the Requesting Rate *RR* to about 12.6922. Joining nodes are discovered relatively quickly, and there are no new leaving nodes to detect because *LR* = 0. Thus, the Membership Accuracy *MA* increases from about 0.8251 to about 0.9739. The Match Probability *MP* still remains high, and hovers around the analytic expectation 0.9817.

In the fourth scenario, the value of the Leaving Rate *LR* is high and the value of the Joining Rate *JR* is low (*LR* = 300, *JR* = 0). Because the value of *LR* is high, the values of *LND* increase to about 0.2669. The values of *JND* remain low, because *JR* = 0. The Combined Adaptive Membership Protocol increases the Requesting Rate *RR* to about 16.5080, in order to detect leaving nodes more quickly. However, leaving nodes are still not detected quickly enough, so *LND* increases and the Membership Accuracy *MA* decreases to about 0.7307 at the end of the fourth scenario. Finally, the Match Probability *MP* remains high, and hovers around the analytic expectation 0.9817, in contrast to the Non-Adaptive Membership Protocol where the Match Probability *MP* fluctuates and decreases to about 0.9.

Lastly, the fifth scenario has a low value of the Leaving Rate *LR* and a low value of the Joining Rate *JR* (*LR* = *JR* = 0). The Combined Adaptive Membership Protocol decreases the value of the Requesting Rate *RR* to about 5.6349, in order to reduce the message cost. The Membership Accuracy *MA* increases to, and remains at, about 0.9375 during most of the fifth scenario. Moreover, the Match Probability *MP* remains high, and hovers around the analytic expectation 0.9817. Note that, with the Combined Adaptive Membership Protocol, the Match Probability *MP* remains high in all five scenarios despite substantial membership churn and substantial changes in the Leaving Rate *LR* and the Joining Rate *JR*.

Finally, we find the averages for each of the membership accuracy, match probability, response time and message cost over all five scenarios. Table 5 shows the overall values of these metrics for the Non-Adaptive, the Retry *R*, the Adaptive *RR* and the Combined Adaptive Membership Protocols, averaged over all five scenarios. As we see from the table, the Combined Adaptive Membership Protocol achieves a membership accuracy of 0.8982, which is quite good. Moreover, the Combined Adaptive Membership Protocol achieves the best match probability 0.9858 of all four protocols. The response time of the Combined Adaptive Membership Protocol is slightly more than that of the Retry *R* Membership Protocol, and is much more than that of the Non-Adaptive and Adaptive *RR* Membership Protocols. The message cost of the Combined Adaptive Membership Protocol is less than that of the Adaptive *RR* Membership Protocol, and is slightly more than that of the Non-Adaptive and Retry *R* Membership Protocols, as Table 5 shows.

Table 5: Non-Adaptive vs. Retry *R* with *Try* = 2 vs. Adaptive *RR* with *Try* = 1 and *RRMax* = 100 vs. Combined Adaptive with *Try* = 2 and *RRMax* = 50

	Non-Adaptive	Retry *R*	Adaptive *RR*	Combined Adaptive
Membership Accuracy	0.8149	0.8542	0.9217	0.8982
Match Probability	0.9704	0.9841	0.9801	0.9858
Response Time	6.0	10.8262	6.0	11.0339
Message Cost	5.0490	5.7335	8.8284	6.3305

Overall, these experiments demonstrate that the Combined Adaptive Membership Protocol is effective in discovering newly joining nodes and in detecting non-operational (leaving) nodes. When the Joining Rate *JR* and the Leaving Rate *LR* are high, the Combined Adaptive Membership Protocol quickly increases the Requesting Rate *RR* to obtain a high membership accuracy. Moreover, when the Joining Rate *JR* and the Leaving Rate *LR* are low, the Combined Adaptive Membership Protocol decreases the Requesting Rate *RR*, in order to maintain a reasonable response time and a reasonable message cost, while still maintaining a reasonable membership accuracy and a high match probability. As a result, the Combined Adaptive Membership Protocol works well not only when the membership is subject to a lot of churn, but also when the membership is stable.

11 Conclusion

We have presented four membership protocols for the iTrust network, the Non-Adaptive, Retry, Adaptive and Combined Adaptive Membership Protocols. These membership protocols allow each member to maintain its own local view of the membership, and aim to keep that local view close to the actual membership. A node that receives a request sends a response, to the requesting node, that contains newly joined member(s) in its local view. If the keywords in the query match metadata that it holds, the node also sends the URL of the document.

A requesting node discovers newly joining nodes from the responses it receives to its requests. Likewise, a requesting node detects leaving (non-operational) nodes when it does not receive responses from those nodes before a timeout occurs, or when it receives an error code from TCP. Thus, the iTrust membership protocols exploit messages already required by the iTrust messaging protocol for distributing metadata and requests.

As our experiments show, for appropriate values of the parameters, the membership accuracy, the response time and the message cost are reasonable, and the match probability is high, particularly for the Combined Adaptive Membership Protocol. The Combined Adaptive Membership Protocol works well not only when the membership experiences a lot of churn but also when the membership is stable.

ACKNOWLEDGMENTS

This research was supported in part by the U.S. National Science Foundation grant number NSF CNS 10-16193.

REFERENCES

[1]. Badger, C.M., L.E. Moser, P.M. Melliar-Smith, I. Michel Lombera, Y.T. Chuang, *Declustering the iTrust search and retrieval network to increase trustworthiness*. Proceedings of the 8th International Conference on Web Information Systems and Technologies, Porto, Portugal, April 2012: p. 312-322.

[2]. Chandra, T.D., V. Hadzilacos, S. Toueg, B. Charron-Bost, *On the impossibility of group membership*. Proceedings of the 15th ACM Symposium on Principles of Distributed Computing, Philadelphia PA, May 1996: p. 322-330.

[3]. Chockler, G.V., I. Keidar, R. Vitenberg, *Group communication specifications: A comprehensive study*. ACM Computing Surveys, 2001. 33(4): p. 427-469.

[4]. Chuang, Y.T., P.M. Melliar-Smith, L.E. Moser, I. Michel Lombera, *Discovering joining nodes and detecting leaving nodes in the iTrust membership protocol*. Proceedings of the 2013 IAENG International Conference on Computer Science, Hong Kong, China, March 2013: p. 189-194.

[5]. Chuang,Y.T., I. Michel Lombera, L.E. Moser, P.M. Melliar-Smith, *Trustworthy distributed search and retrieval over the Internet.* Proceedings of the International Conference on Internet Computing, Las Vegas, NV, July 2011: p. 169-175.

[6]. Clarke, I., O. Sandberg, B. Wiley, T. Hong, *Freenet: A distributed anonymous information storage and retrieval system.* Proceedings of the Workshop on Design Issues in Anonymity and Unobservability, Berkeley, CA, July 2001: p. 46-66.

[7]. Cuenca-Acuna, F.M., C.Peery, R.P. Martin, T.D. Nguyen, *PlanetP: Using gossiping to build content addressable peer-to-peer information sharing communities*. Proceedings of the 12th IEEE International Symposium on High Performance Distributed Computing, Seattle, WA, June 2003: p. 236-246.

[8]. Freedman, M.J. and R. Morris, *Tarzan: A peer-to-peer anonymizing network layer*. Proceedings of the 9th ACM Conference on Computer and Communications Security, Scottsdale, AZ, November 2014: p. 193-206.

[9]. Ganesh, A., A.M. Kermarrec, L. Massoulie, *Peer-to-peer membership management for gossip-based protocols.* IEEE Transactions on Computers, February 2003. 52(2) : p. 139-149.

[10]. Gnutella, http://en.wikipedia.org/wiki/Gnutella (2000).

[11]. Gramoli, V., A.M. Kermarrec, E. Le Merrer, *Distributed churn measurement in arbitrary networks.* Proceedings of the 27th ACM Symposium on Principles of Distributed Computing, Toronto, Canada, August 2008: p. 431.

[12]. Leng, C., W.W. Terpstra, B. Kemme, W. Stannat, A.P. Buchmann, *Maintaining replicas in unstructured P2P systems.* Proceedings of the ACM Conference on Emerging Networking Experiments and Technologies, Madrid, Spain, December 2008: p. 19.

[13]. Liu, H., X. Liu, W. Song, W. Wen, *An age-based membership protocol against strong churn in unstructured P2P networks.* Proceedings of the 2011 International Conference on Network Computing and Information Security, vol. 2, Guilin, China, May 2011: p. 195-200.

[14]. Lv, Q., P. Cao, E. Cohen, R. Li, S. Shenker, *Search and replication in unstructured peer-to-peer networks.* Proceedings of the 16th International Conference on Supercomputing, Baltimore, MD, June 2002: p. 84-95.

[15]. Melliar-Smith, P.M., L.E. Moser, I. Michel Lombera, Y.T. Chuang, *iTrust: Trustworthy information publication, search and retrieval*. Proceedings of the 13th International Conference on Distributed Computing and Networking, LNCS 7129, Hong Kong, China, January 2012: p. 351-366.

[16]. Michel Lombera, I., Y.T. Chuang, P.M. Melliar-Smith, L.E. Moser, *Trustworthy distribution and retrieval of information over HTTP and the Internet*. Proceedings of the 3rd International Conference on the Evolving Internet, Luxembourg City, Luxembourg, June 2011: p. 7-13.

[17]. Mischke, J., and B. Stiller, *A methodology for the design of distributed search in P2P middleware*, IEEE Network, 2004. 18 (1): p. 30-37.

[18]. Peng, P., L.E. Moser, P.M. Melliar-Smith, Y.T. Chuang, I. Michel Lombera, *A distributed ranking algorithm for the iTrust information search and retrieval system*. Proceedings of the 9th International Conference on Web Information Systems and Technologies, Aachen, Germany, May 2013: p. 199-208.

[19]. Pruteanu, A., V. Iyer, S. Dulman, *ChurnDetect: A gossip-based churn estimator for large-scale dynamic networks*. Proceedings of the Euro-Par 2011 Conference, LNCS 7155, Bordeaux, France, August 2011: p. 289-301.

[20]. Reiter, M.K.. and A.V. Rubin, *Crowds: Anonymity for Web transactions*. ACM Transactions on Information and System Security, November 1998. 2 (1): p. 66-92.

[21]. Richardson, S. and I.J. Cox, *Estimating global statistics for unstructured P2P search in the presence of adversarial peers.* Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, Gold Coast, Queensland, Australia, July 2014: p. 203-212.

[22]. Schiper, A. and S. Toueg, *From set membership to group membership: A separation of concerns*. IEEE Transactions on Dependable and Secure Computing, 2006. 3(1): p. 2-12.

[23]. Terpstra, W.W., J. Kangasharju, C. Leng, A.P. Buchmann, *Bubblestorm: Resilient, probabilistic, and exhaustive peer-to-peer search*. Proceedings of the ACM Conference on Applications, Technologies, Architectures and Protocols for Computer Communications, Kyoto, Japan, August 2007: p. 49-60.

[24]. Voulgaris, S., D. Gavidia, M. Van Steen, *CYCLON: Inexpensive membership management for unstructured P2P overlays.* Journal of Network and Systems Management, June 2005. 13(2): p. 197-217.

[25]. Zage, D., C. Livadas, E.M. Schooler, *A network-aware distributed membership protocol for collaborative defense*. Proceedings of the International Conference on Computational Science and Engineering, vol. 4, 2009: p. 1123-1130.

TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

Do Personal Attributes and An Understanding of Sarcasm and Metaphor Explain Problematic Experiences on the Internet? —A Survey for the Development of Information Literacy Education Tools—

[1]Yuhiko Toyoda, [2]Mika Takeuchi, [3]Hiroshi Ichikawa, [4]Mitsuteru Tashiro and [5]Masao Suzuki
[1]School of Business administration, Sanno University, Japan;
[2]Department of Humanities and Social Sciences, Jissen Women's University, Japan;
[3]Department of Home Economics, Otsuma Women's University, Japan;
[4]Institute for InfoSocionomics, Tama University, Japan;
[5]Department of Human Sciences, Waseda University, Japan;
toyoda_yuhiko@hj.sanno.ac.jp; takeuchi-mika@jissen.ac.jp; ichikawa.h@otsuma.ac.jp;
tashiro@ni.tama.ac.jp; masaosuz@waseda.jp

ABSTRACT

The Internet today provides users with a great amount of convenience due to the improvement of tools and their functions, and the expansion in the numbers of users. However, with the expansion of both online content and time spent online, new and potential risks have emerged in this virtual space. A majority of Japanese students experience communication troubles over the Internet, and therefore higher education institutions have increased their efforts to reduce risks on the Internet by, for example, offering basic security education through information literacy programs. However, because of the number and variability of Internet risks, it is unlikely that these efforts have achieved satisfactory results. As one of the risk factors inducing troubles on the Internet, where users exchange information with other (anonymous) users, we examine the immaturity of the "theory of mind," which is generally used to infer the conditions, viewpoints, and comprehension level of others.

This study reports the results of a self-report questionnaire used to examine the problems or difficulties encountered on the Internet by young women in Japan. It focuses on the relationship between the comprehension of sarcasm and metaphor expressions, provided for the purpose of estimating the "theory of mind," as well as of Internet terminology and the problematic experiences. The problems identified were roughly classified into four different categories: addictive Internet use, dishonest/illegal dealings, communication gaps, and shopping-related difficulties. Multiple regression analyses was conducted, aimed at identifying factors that could explain such problems. The results suggested that personal attributes and sensation-seeking tendencies, including social vulnerability and anxiety, and the understanding of sarcasm, metaphor expressions, and Internet terminology, might be significantly correlated with the problems experienced. On the basis of this study, proposals are made as to what areas should be focused on in information literacy education programs in the future.

Keywords: Internet Risk, comprehension of metaphor and sarcasm, theory of mind, educational tool

DOI: 10.14738/tnc.32.1169
Publication Date: 02th May, 2015
URL: http://dx.doi.org/10.14738/tnc.32.1169

1 Introduction

As the Internet has become more ubiquitous in both private and public life, the access to information has drastically improved. However, at the same time the risks related to Internet use have also grown. With these growing risks in mind, the Organization for Economic Co-operation and Development (OECD) published a recommendation in 2012 on the risk of Internet use by adolescents, entitled "The Protection of Children Online[1]. Recommendation of the OECD Council, Report on risks faced by children online and policies to protect them." Following these OECD recommendations, Japan's Ministry of Internal Affairs and Communications published a summary report on the Internet Literacy Assessment indicator for Students (ILAS)[2]. This report found that children who had received Internet security education or had discussed it with their family were much more aware of the security concerns. In line with efforts to increase awareness, this study presents the results of a survey on female students' personal attributes and their understanding of sarcasm, metaphor, and Internet terminology—all of which were speculated to be related to difficulties experienced on the Internet—and also examines the predictability of these factors.

2 Background and Literature Review

Electronic communication media such as Twitter, Facebook, and Line are extensively used in Japan, especially by adolescents. When using these communication methods, although it is easy to transmit messages, a simple error or a poorly chosen comment may lead to difficulties. For example, in recent years, a convenience store was forced to close because one of its employees posted a photo of himself lying in a freezer to provide a topic of conversation for his friends. In another incident, a student posted a photo of a stranger online accompanied with a defamatory comment about the stranger; as a result his university was flooded with complaint calls and he was eventually forced to drop out of school.

2.1 The difference between "Enjoh" and "Flaming"

Hirai [3] found that the origin of the so-called "enjoh" (a Japanese term literally meaning "fire spreading") was an incident on a blog where abusive and slanderous comments were made about a comment posted by an ordinary user, and then examined several similar, widespread cases. Hirai cited Thurlow et al., [4] to differentiate between "flaming" and "enjoh." In his study, flaming was defined as a "hostile and aggressive mutual act in computer-mediated communication (Thurlow et al.,). "Enjoh" in Japan is defined by Hirai [3] as a "phenomenon in which a message contributed to such services as blogs, mixi or Twitter, as well as the person who posted such a message, [is] flooded with criticism and blame." Interestingly, "at first, 'enjoh' referred to only some of the events or topics in the electronic forum named '2channel'." O'Sullivan and Flanigan [5] described "enjoh" and "flaming" as separate phenomena by examining the differences in those involved, saying "even if there is an exchange of words which seem to be denouncing (others) online, the norm of the mutual act is not violated and such a case will not be considered flaming, which is a hostile and aggressive mutual act, in case those who exchange messages recognize that what they are doing is a 'word game' or a 'joke'."

Twitter is a highly representative cyber-community in which specific users participate. The characteristics of the user communities have been described in a number of studies (Takeichi & Sasahara et al., [6]; Kwak et al., [7]; Bollen et al., [8]; Grabowicz et al., [9]; Conover et al., [10]). Highly particular network structures between users (Kwak et al., [7]; Bollen et al., [2]), particularity in the nature of social negotiation (Grabowicz et al., [9]; Conover et al., [10]), information diffusion (Romero et al., [11]; Weng et al., [12]), collective attention (Lehmann et al., [13]), collective mood

(Golder and Macy, [14]), and user dynamics related to particular real-life events (Sakaki et al., [15]; Borge-Holthoefer et al., [16]; Gonz´alez-Bail´on et al., [17]) have all been examined. Twitter, as a simultaneous communication medium, is used by a community of users to search for events and topics that are happening now (Takahashi et al., [18]), and is even referred to for trend searches in the stock markets (Bollen et al., [19]). It has been described as a "social sensor" (Takeichi & Sasahara et al., [6]). The Great Eastern Japan Earthquake drew attention to the intricate functions of Twitter, and the use of the service to search for chronological traces or "digital fossils," by looking for events that received collective attention and collective responses, was examined in detail (Sasahara, K., Hirata, Y., Toyoda, M., Kitsuregawa, M. & Aihara, K., [20]).

The boundary between "enjoh" and "flaming" becomes more ambiguous as the online community, which once consisted of sensible "experts" who were able to distinguish between "jokes among friends" and serious violations of fundamental human rights, grows to include all people in society.

Because anybody can easily participate in cyber space anonymously, the ambivalence of the vulnerability caused by a lack of strong protection of personal information remains unresolved.

2.2 Human errors, "problems" and "potential problems" in cyber communication

Serious social problems are often caused by releasing information without understanding the nature of the Internet. Many senders of problematic messages are often not interested in the security of their own personal information, and the information left in their SNS accounts or blogs is readily searched for and collected. It is not rare for the name of their school and their residential address to be disclosed on the web. As a result, some institutions have had their operations hindered or suspended, and there have been tragic cases of individuals forced to quit their schools or jobs.

Heinrich's law states that behind great accidents there are smaller cases in the preparatory stage. Heinrich [21], the father of the safety study, was the first to introduce the rule of thumb according to which "if a major accident occurs, 29 medium-scale accidents have occurred in the surroundings and 300 small-scale accidents behind." For major incidents related to Internet communication, in reference to this general principle of accident potency, it is likely that the number of problems that actually occur is much more than the number of cases recognized as problems. In information literacy education, it may therefore be necessary to exclude estimations based on wishful thinking as to the number of "potential problems."

The "error of prediction" refers to human errors that cause accidents. Errors of prediction occur in cyber space as a result of the action selected by a transmitter who does not sufficiently predict public reaction to the information or who remains incapable of careful estimation. To prevent errors, it is essential to improve the precision of the predicted result of the action; that is, to collect accurate information in advance so as to make adaptive choices.

One of the features of cyber space communication is that the information sender and receiver do not share a time and space face-to-face, and that the participants are not necessarily specified. This makes it difficult to perform the basic safety measures of prediction and control. In face-to-face communication, it is expected that a person's point of view is different from others and that there is a gap in the quality and quantity of information a person has, which means that it is not easy for two or more people to arrive at a mutually sufficient level of "comprehension" and satisfaction. The most

central tasks for human social development include being able to reach a common viewpoint (by trial and error) that encompasses the differences between two people, and the process of learning the skills needed to coordinate with others. Cyber communication imposes a learning task on users that is more difficult to solve than a task related to direct communication.

2.3 Possibility of applying the "theory of mind" to cyber communication

The "theory of mind" is the overall process of estimating other people's viewpoints (by abandoning one's own), adequately identifying the difference between the information used by others and the information belonging to oneself, and then coordinating and selecting one's own behavior. Studies on the "theory of mind" encompass ethology, pedagogy, cognitive science, psychology and clinical developmental psychology, especially in the field of the rehabilitation of children and adolescents with pervasive developmental orders (PDD, especially Asperger's Syndrome or autism spectrum disorders) according to the needs and interests in each field. Research in this area was first conducted by Premack & Woodruff [22], who observed the behavior of primates such as chimpanzees. Premack [23] considered that if animals or human beings could understand what they themselves or others intend, know, believe, think, doubt, suppose, pretend or like, then they could be regarded as having a "theory of mind."

Hobson [24] and Moore [25] stated that "a child tries to understand others as those who are similar to him/herself in that they are sources of subjective attitudes but sometimes have a mental attitude that is different from his/hers, as individual beings that have peculiar attitudes and viewpoints." Hobson and Moore believed that children formed the basic framework for an understanding of the "self and others" when they went through such a process. Baron-Cohen et al [26] considered it an important goal of the "theory of mind" to differentiate between those who have a correct belief by themselves and noticing those who have a wrong belief, and examined the developmental process using a false-belief task.

The "Sally-Anne Test" is typical of a false-belief task. This task consists of asking participants (children) where Sally (an "other" or distinct person), whose doll has been moved somewhere else by someone (Anne) during Sally's absence, could look for the doll. In this task, the "theory of mind" is necessary for participants to able to notice the difference between having a correct belief (i.e., they know that the doll is not at the original place) and being aware that the leading character of the story (Sally) has an incorrect belief (i.e., that the doll is in the original place). This task, also referred to as an "unexpected task," requires participants to abandon their own beliefs and viewpoints and to put themselves in the position of others.

There have been debates as to the point in the developmental stage at which this psychological process commences or is achieved. Research on this theme has developed along various paths, and has been included in brain function development trials. In recent years, functional neuroimaging studies have been actively conducted using fMRI (function MRI) (for example, see Takamiya, Matsui, Kobayashi et al, [27] and Moriguchi, [28]). During the execution of a "theory of mind" task with normal adolescents, when the task was accomplished, an activation of the medial and lateral prefrontal cortexes and the orbito-frontal cortex was observed. A study on the personality attributes related to psychosomatic disorders indicated that alexithymia (difficulty in identifying and describing emotions in the self) was characterized by self-emotional disturbances which included difficulty in realizing a person's sentiments in the act of representation. Moriguchi reported that changes in the function of the medial prefrontal cortex had been observed, which helped to represent a person's own mind and an understanding of others.

Adachi et al., [29] conducted an experiment aimed at investigating the situation recognition skills and related particularities of participants with attention deficit/hyperactivity disorder (AD/HD), Asperger's Syndrome (AS) or pervasive developmental disorders including high functioning autism, using a metaphor and sarcasm scenario test (MSST). It was found that participants with AS had difficulty understanding sarcasm even though their language ability was good, while there was no difference between the scores for metaphor and sarcasm in the group with high functioning autism (HFA) and groups with AD/HD. Adachi (2006) pointed out that "sarcasm contains negative meanings that are contrary to the wording, such as blame and ridicule. Sarcasm is valid only when it has been agreed between the speaker and the receiver of the sarcasm that what is said is contrary to the truth. That is to say, the ability to understand the implied meaning is based on judgment abilities which are different in quality for metaphoric sentences and sarcastic sentences, namely a judgment of the fact for the former and judgment of the value for the latter." It was found that the value judgment ability necessary to share sarcasm was less dependent on language ability than the ability to share metaphorical sentences, as the understanding of sarcasm does not depend on age and language ability while the understanding of metaphor does.

The "theory of mind" includes a meta-representation of oneself and others. A series of "unexpected tasks" and "sarcasm-understanding tasks," which have been developed as functional development and achievement in human socialization indexes, contain stories with unusual settings. The tasks can be enlarged to include daily scenes, so that the execution level can be estimated. Participants are required to depart from their own linear beliefs and viewpoints to carry out a meta-representation task which consists of readily moving those representations onto others.

Today, young people spend a great deal of their life interacting in cyber space. It is necessary to examine how these young people, who are living daily examples of "enjoh," "flaming," or fraud, can build meta-representations in their relationships with themselves and others in cyber space. Communication in cyber space has unknown processes that are still considered to be in the experimental stage. It may be possible that there are incidents that have been caused by an incomplete "theory of mind" in terms of the "problems" that have occurred and the "potential problems" that could occur. This speculation suggests that the content of current information literacy education needs to be re-examined and expanded.

This study examines the results of a self-report questionnaire completed by university students to obtain basic data on their experience of "problems" and "potential problems" in their use of cyber space, as well as on the personal attributes related to such experiences and the achievement of the "theory of mind."

3 Methods

3.1 Participants

The survey was conducted on 454 female students studying at a junior college at two women's universities in Tokyo prefecture, using a self-report format.

These respondents were all attending a commuting course, in which the researchers are responsible for information literacy education. Respondents agreed on the significance and necessity of the epidemiological research, having fully understood the necessity of information literacy through lectures at the university. This survey was conducted with educational urgency, as it was felt that

female adolescents suffer larger psychosocial, psychosomatic and mental damage than their male colleagues if they meet with accidents in cyber space. However, when the questionnaire was distributed, it was explained that participation was voluntary, that respondents could leave blank the questions they did not wish to answer, that they could stop answering questions when they wished, that they did not have to fill in their names, and that the data would be mechanically treated by a trained company with a duty of confidentiality. There is no conflict of interest in this study.

3.2 Period of the survey

We conducted the survey in July 2014 in a lecture room at the university. At the end of the lecture, each student was asked to respond once to the survey items on a voluntary basis.

3.3 Research battery

The questionnaire consisted of 60 Risk Management Test (RMT) questions developed by Takeuchi and Suzuki [30] for education against fraudulent sales, as well as questions covering the use of the Internet, problems experienced on the web, and multiple-choice interpretative questions regarding sarcasm and metaphors and knowledge of Internet terminology. The questions were organized as follows:

a. A group of questions related to the use of cyber communication, including devices used for connection and SNS sites accessed.
b. A Risk Management test (RMT): This is a self-evaluation sheet for crisis management education, and has groups of questions related to “physiological needs,” “safety needs,” “social belonging needs,” “esteem needs,” and “self-actualization needs,” based on Maslow’s principle of the hierarchy of needs [31]. There is also a group of questions linked to “sensation-seeking and curiosity,” built on a model proposed by Cloninger [32]. The responses were given on a scale of one to five, with 1 = “not at all” and 5 = “always.”
c. Problematic experiences related to the use of the Internet (IT): Items identified in a previous study by Tashiro [33] were described and respondents were asked whether they had experienced similar cases. The responses were from 0 = “never” to 4 = “very often.”
d. Metaphor and Sarcasm Scenario Test-for Adolescents (MSST-A) and the Internet: To test understanding of sarcasm and metaphors, we developed questions that focused on daily scenes familiar to the young respondents, with choices related to the understanding of these scenes and the related behaviors. Using the same format, we developed questions to check respondent understanding by citing examples of conversations between friends and the use of “Internet terminology.”
e. Internet Terminology Questionnaires-2: The “terminology” used in cyber communication is unique. Often such specialized jargon is common only to the members of the community concerned. “Internet terminology” changes regularly and there are many terms in daily use that have originated from cyber space. As the level of knowledge and use of “Internet terminology” is expected to play a role in an index which measured the degree of cyber communication penetration in a respondents’ life, we decided to identify typical “Internet terms” that are often seen but not so widely used in real-world conversations. In this survey, we asked respondents to answer the questions using five scales from 0 = “I have never seen it” to 4 = “I use it frequently.”

4 Results

4.1 Problems experienced on the Internet (IT)

Only 14 out of the 452 valid responses (3.1%) said that they had experienced "no trouble at all" related to Internet use. For the remainder, many said that these problematic experiences were related to their exposure to uncertainty, danger and unpleasant information as information receivers, such as "I have believed false rumors posted on the Internet," "I have received inviting mails/messages from strangers," and "I have received disgusting mails/messages." The experience rate for situations which indicated addictive Internet access, such as "I spend so much time on the Internet that I think I have difficulty in daily life," was also high. In addition, 46.8% of respondents indicated that they had experienced problems as senders of information, such as "I have disgusted others by mails/messages that I sent."

4.2 Factor analysis for the problematic experiences on the Internet

We conducted a factor analysis to examine the main patterns for problematic experiences on the Internet. Table 1 shows the result of the maximum likelihood estimation and the promax rotation. For the first factor, the factor loading for the items "I spend so much time on the Internet that I think I have difficulty in daily life," and "I spend so much time playing games that I think I have difficulty in daily life" was high, showing that this factor was related to "addiction tendency" (contribution ratio = 12.817%, $\alpha = 0.579$). For the second factor, items related to dishonest/illegal experiences, such as "I have posted false rumors on the Internet" and "I have accessed a computer belonging to another person before" were extracted (contribution ratio = 16.366%, $\alpha = 0.660$). For the third factor, communication-related items, including "I have received inviting mails/messages from strangers" and "I have received disgusting mails/messages" were extracted (contribution ratio = 12.671%, $\alpha = 0.713$). For the fourth factor, shopping-related items such as "Some of the articles that I ordered on Internet-order or auction services on the Internet have not arrived yet" and "Some of my requests for return have not been accepted by Internet-order or auction services" were extracted (contribution ratio = 7.676%, $\alpha = 0.736$). In a later analysis, we calculated the synthesized score for each factor, which was obtained by summing up the item raw scores from which the factor had been extracted.

4.3 Factor analysis for the Risk Management Test (RMT)

The RMT is a questionnaire sheet prepared for use in educational programs to allow students to raise their awareness by self-evaluating the risks inherent in their personal attributes that could induce fraud or solicitation (Takeuchi & Suzuki, 2000). To examine the image structures, we conducted a factor analysis using maximum likelihood estimation and promax rotation (Table 2). When the first five factors were converted, the cumulative contribution ratio was 29.9%. For the first factor, 11 items were extracted, including "I am sensitive to fashion for my clothes, hairstyle and makeup," "I adore the truth and perfect beauty and I hope to approach such things by myself," "I want to have friends and a boyfriend I can boast of to others," and "I love top-brand articles, even if they are expensive." Although the items cited for the first factor seem to be diverse at first glance, they all include items that refer to the possibility of high suggestibility, anxiety and evaluation by others. In summary, this can be understood as a factor which shows the "tendency to obey a person silently (according to an influence from an external environment)" (contribution ratio = 10.28%, Cronbach's $\alpha = .714$). For the second factor, items showing "sensation-seeking," such as "I love to

feel thrilled riding jet coasters and trying free falls in amusement parks" and "I want to try bungee jumping and sky diving" (contribution ratio = 7.02%, Cronbach's α = .579), were extracted. For the third factor, extracted items related to "anxiety over relationships," such as "I sometimes cannot say what I want for fear of hurting others" or "I tend to be unable to say no when asked to do something by others" (contribution ratio = 3.68%, Cronbach's α = .641). Items associated with the "desire to surpass existing regulations" such as "I hope to complete something 'original' that only I can do" and "There is something I would like to try if I were not punished by law" were extracted for the fourth factor (contribution ratio = 4.30%, Cronbach's α = .390). Finally, for the fifth factor, items linked to the "avoidance of physical pain," including "I want to avoid things that cause pain to my body" and "I do not want to carry out activities that put a physical burden on me" were extracted (contribution ratio = 4.03%, Cronbach's α = .646).

Table 1: Factor analysis for problematic experiences on the Internet

Experienced Internet Problems	Factor Loadings			
	Addiction tendency	Illegality experience	Communication	Shopping
	factor_1	factor_2	factor_3	factor_4
IT12 I spend so much time on the Internet that I have difficulty in daily life.	**.999**	-.003	-.001	.000
IT13 I spend so much time playing games that I have difficulty in daily life.	**.488**	.225	.029	.108
IT10 I have believed false rumors posted on the Internet.	**.321**	.210	.238	-.125
IT09 I have posted false rumors on the Internet.	.171	**.924**	-.173	-.074
IT11 I have accessed a computer belonging to another person without permission.	.124	**.562**	.075	.072
IT08 I have sent mails or made contributions to the Internet including information that should be kept secret.	.274	**.533**	.051	-.050
IT04 I have received mails or messages inviting me to buy such things as illegal drugs.	.103	**.382**	.321	.218
IT05 I have received inviting mails/messages from strangers.	.200	.201	**.731**	-.287
IT06 I have received disgusting mails/messages.	.234	.247	**.604**	-.255
IT03 I have received mails/messages claiming payment that I have nothing to do with.	.165	.222	**.574**	-.119
IT07 I have disgusted others by mails/messages that I sent.	.197	.201	**.206**	.026
IT01 Some of the articles that I ordered on Internet-order or auction services on the Internet have not arrived yet.	.085	.341	.306	**.627**
IT02 Some of my requests for return have not been accepted by the Internet-order or auction services.	.168	.356	.289	**.595**
Contribution ratio(%)	12.817	16.366	12.671	7.676
Cumulative contribution ratio (%)	12.817	29.183	41.854	49.530
Cronbach's α	.579	.660	.713	.736

Table 2: Risk Management Test (RMT) factor analysis

Risk Management Items	Tendency to obey a person silently	Sensation seeking tendency	Anxiety over relationship	Desire to surpass regulations	Avoidance of physical pain
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Q39 I am sensitive to fashion for my clothes, hairstyle and makeup.	**.689**	.010	-.488	-.221	.015
Q29 I adore the truth and perfect beauty and I hope to approach such things by myself.	**.577**	.093	-.049	.081	-.109
Q28 I want to have friends and a boyfriend I can boast of to others.	**.521**	-.003	.051	.016	-.075
Q16 I love top-brand articles, even if they are expensive.	**.491**	.080	-.185	.043	.108
Q03 I feel anxious when I am alone.	**.405**	-.076	.104	.038	.052
Q44 I believe in the charm or the incantation of safety.	**.391**	.076	.001	-.075	-.093
Q23 I feel attracted by pure ways of life.	**.355**	-.006	.203	.093	-.095
Q57 I sometimes change my judgment according to others' opinion.	**.350**	-.134	.273	-.182	.034
Q37 I may be somewhat selfish in order to feel comfortable.	**.340**	.002	.007	.137	.046
Q43 I have given up something that I absolutely wanted to do because of the lack of physical power.	**.338**	.075	.124	.278	.174
Q53 I want to do anything that pleases others.	**.337**	.080	.027	.041	-.288
Q54 I love to feel thrilled riding jet coasters and trying free falls in amusement parks.	.035	**.637**	.192	-.390	.338
Q48 I want to try bungee jumping and sky diving.	-.037	**.631**	.151	-.226	.125
Q18 I like dangerous things and thrilling games.	.083	**.585**	.058	.250	-.052
Q24 I am rather curious.	.215	**.385**	.003	.175	-.155
Q60 I have intentionally tried things that I know are dangerous or illegal.	.150	**.336**	.076	.260	.078
Q06 I am looking for sensations in life.	.242	**.269**	.089	.251	-.044
Q21 I sometimes cannot say what I want for fear of hurting others.	.408	-.295	**.580**	-.242	-.130
Q33 I tend to be unable to say no when asked to do something by others	.272	-.106	**.399**	-.262	-.179
Q17 I hope to complete something "original" that only I can do.	.287	.270	.187	**.458**	-.191
Q42 There is something I would like to try if I were not punished by the law.	.187	.321	.181	**.323**	.032
Q13 I want to avoid things that cause pain to my body.	.275	-.328	.085	.194	**.538**
Q55 I do not want to carry out activities that put a physical burden on me.	.131	-.351	.107	.147	**.482**
Q19 I do not have so much endurance.	.089	-.197	.147	.165	**.349**
Q01 I cannot endure "physically hard" work or tasks.	.269	-.183	.016	.204	**.347**
% of diffusion	10.283	7.024	3.667	4.300	4.025
Cumulative %	10.283	17.307	20.973	25.274	29.299
Cronbach's α	.714	.579	.641	.390	.646

For later analysis, we calculated the synthesized score for each factor, by summing up the raw scores of items from which the factor was extracted.

4.4 Calculation of the Internet terminology comprehension score

We asked respondents to select one out of six choices, including the answer "I do not know," to examine their comprehension of bilateral conversations where sarcasm, metaphor or Internet terminology was used. In real life scenarios, comprehension of sarcasm, metaphor and Internet terminology is in fact extensive and dependent on the context. The goal of this study was to check if those who tended to make choices that deviated from the common comprehension of the majority of dialogue participants might enter into conflicts with others in cyber space communication, where

language is used as the principal medium. As there is no one correct way for the use of sarcasm and metaphor in ordinary dialogue, we calculated the MSST-A score for each respondent by extracting five answers where the selection rate by respondents was between 60 and 70 percent, and then summing up the raw scores for the five questions by granting one point to those who selected the most frequent answer and zero to those who selected other answers. The MSST-A score was applied to later calculations to develop a sarcasm and metaphor comprehension index. Table 3 shows the five MSSA-related questions for the comprehension of Internet terminology.

4.5 Multiple regression analysis using problematic Internet experiences as dependent variables

a) Multiple regression analysis using problematic Internet experiences as the dependent variables and the RMT synthesized factor scores as independent variables:

For each of the four problematic Internet experience categories classified through the exploratory factor analysis, a multiple regression analysis was conducted using the forced entry method, so as to show the degree of explanation ratio for the RMT synthesized factor scores as personal attributes. It was found that the standardized R2 = .074 (Pr. = .000) for the item, "addiction tendency (to the Internet) (IT_1)" was significant. Of the personal attribute variables (RMT) entered for "addiction tendency," "avoidance of physical pain: RMTscr_5" (β = .193, Pr. = .000) and the "tendency to obey a person silently (according to an influence from the external environment): RMTscr_1" (β = .124, Pr. = .015) were found to be especially significant descriptive factors. The standardized R2 = .024 (Pr. = .007) for "[dishonest/]illegal experiences (IT_2)" was also shown to be significant. Of the personal attribute variables (RMT) entered for "[dishonest/] illegal experiences," "sensation seeking tendency: RMTscr_2" (β = .128, Pr. = .015) was found to be an especially significant descriptive factor. The standardized R2 = .041 (Pr. = .000) for "communication (IT_3)" was found to be significant. For the personal attribute variables (RMT) entered for "communication," "sensation seeking tendency: RMTscr_2" (β = .138, Pr. = .008) and the "desire to surpass existing regulations: RMTscr_4" (β = .013, Pr. = .049) were found to be especially significant descriptive factors. It was also found that the standardized R2 = .028 (Pr. = .027) for "shopping (IT_4)" was significant. For the personal attribute variables (RMT) entered for "shopping," "desire to surpass existing regulations: RMTscr_4" (β = .20, Pr. = .016) was found to be an especially significant descriptive factor. Although these factor values are low, we can see that there are Internet usage behavior patterns that could induce problems, as the different personal attribute factors suggested significant explicability for each problematic Internet experience (see Table 4).

For later analysis, we calculated the synthesized score for each factor, by summing up the raw scores of items from which the factor was extracted.

4.6 Calculation of the Internet Terminology Comprehension Score

We asked respondents to select one out of six choices, including the answer "I do not know," to examine their comprehension of bilateral conversations where sarcasm, metaphor or Internet terminology was used. In real life scenarios, comprehension of sarcasm, metaphor and Internet terminology is in fact extensive and dependent on the context. The goal of this study was to check if those who tended to make choices that deviated from the common comprehension of the majority of dialogue participants might enter into conflicts with others in cyber space communication, where language is used as the principal medium. As there is no one correct way for the use of sarcasm and metaphor in ordinary dialogue, we calculated the MSST-A score for each respondent by extracting five answers where the selection rate by respondents was between 60 and 70 percent, and then

summing up the raw scores for the five questions by granting one point to those who selected the most frequent answer and zero to those who selected other answers. The MSST-A score was applied to later calculations to develop a sarcasm and metaphor comprehension index. Table 3 shows the five MSSA-related questions for the comprehension of Internet terminology.

4.7 Multiple regression analysis using problematic Internet experiences as dependent variables

a) Multiple regression analysis using problematic Internet experiences as the dependent variables and the RMT synthesized factor scores as independent variables:

For each of the four problematic Internet experience categories classified through the exploratory factor analysis, a multiple regression analysis was conducted using the forced entry method, so as to show the degree of explanation ratio for the RMT synthesized factor scores as personal attributes. It was found that the standardized R2 = .074 (Pr. = .000) for the item, “addiction tendency (to the Internet) (IT_1)” was significant. Of the personal attribute variables (RMT) entered for “addiction tendency,” “avoidance of physical pain: RMTscr_5” (β = .193, Pr. = .000) and the “tendency to obey a person silently (according to an influence from the external environment): RMTscr_1” (β = .124, Pr. = .015) were found to be especially significant descriptive factors. The standardized R2 = .024 (Pr. = .007) for “[dishonest/]illegal experiences (IT_2)” was also shown to be significant. Of the personal attribute variables (RMT) entered for “[dishonest/] illegal experiences,” “sensation seeking tendency: RMTscr_2” (β = .128, Pr. = .015) was found to be an especially significant descriptive factor. The standardized R2 = .041 (Pr. = .000) for “communication (IT_3)” was found to be significant. For the personal attribute variables (RMT) entered for “communication,” “sensation seeking tendency: RMTscr_2” (β = .138, Pr. = .008) and the “desire to surpass existing regulations: RMTscr_4” (β = .013, Pr. = .049) were found to be especially significant descriptive factors. It was also found that the standardized R2 = .028 (Pr. = .027) for “shopping (IT_4)” was significant. For the personal attribute variables (RMT) entered for “shopping,” “desire to surpass existing regulations: RMTscr_4” (β = .20, Pr. = .016) was found to be an especially significant descriptive factor. Although these factor values are low, we can see that there are Internet usage behavior patterns that could induce problems, as the different personal attribute factors suggested significant explicability for each problematic Internet experience (see Table 4).

Table 3 Metaphor and Sarcasm Scenario Test-for Adolescent (MSST-A) and the response rate of items applied to the Internet terminology comprehension scores (N = 450)

MSST-A Question	Answer	%
30. When I eat all the snacks that are left at home, my younger brother says, "You have a good character." What does he want to say?	1 I think he praises my character.	2.2%
	2 He says that I have eaten all the snacks.	6.0%
	3 He is afraid if I will have a stomachache.	0.2%
	4 He wants to eat them too.	21.6%
	5 I should have left some for other people.	**65.6%**
	6 I do not know.	4.7%
33. When I am putting in order clippings featuring my favorite artist, my elder sister says: "You are studying eagerly." What does she want to say?	1 She just says the fact that I am studying.	16.7%
	2 She points out that I am eager for things I like although I do not study other things.	**60.7%**
	3 I think she is afraid if I will be tired.	0.7%
	4 She says that because she also wants to be enthusiastic about something.	11.6%
	5 She says that I am doing "something good."	5.6%
	6 I do not know.	4.9%
38. I frame a tweet saying that I have found a part-time job with high hourly pay and my friend replies to me by saying "Information kwsk (Internet slang for "tell me more")." What is he thinking of?	1 He thanks me for getting information about the job.	12.4%
	2 He wants me to modify the wrong information that I tweeted.	0.4%
	3 He thinks that I do too much part-time job instead of doing what I should do.	0.4%
	4 He is also interested in this job.	**69.6%**
	5 His is not well paid for his job.	0.2%
	6 I do not know.	17.1%
42. On the occasion of a scandal involving a politician, I check the related Internet forum and find a contribution that a user made saying "He will resign soon, meshiuma (literally, "my meal will be delicious!")." What is his/her feeling?	1 He/she feels sorry that the politician will resign.	1.8%
	2 He/she believes that politicians should behave correctly.	7.8%
	3 He/she is happy with someone else's unhappiness.	**69.1%**
	4 He/she hopes that political reform will continue.	3.3%
	5 The person who makes this comment knows that he/she is having a good meal.	1.3%
	6 I do not know.	16.2%
43. When I make a proposition to hold a voluntary study meeting as the lecture is too difficult, my classmate says "Sore Daretoku? (Who will benefit from this?)." What is his/her feeling?	1 He/she is disappointed that the lecture is too difficult.	1.1%
	2 "It is too troublesome. It is rare for you to act goody-goody like this."	17.6%
	3 "I do not want to do troublesome things. This will be no good for anyone."	**65.1%**
	4 He/she thinks that this is a useful proposition for everyone.	2.0%
	5 He/she thinks that this will be a good opportunity for him/her to achieve a good score.	1.3%
	6 I do not know.	12.7%

Table 4 Multiple regression analysis using the total score for various problematic Internet experience factors (IT_1-IT_4) as the dependent variables and the RMT synthesized factor scores (RMTscr_1- RMTscr_5) as the independent variables (using the forced entry method)

	IT_1 Addiction tendency Standardized β	IT_2 Illegality experience Standardized β	IT_3 Communication Standardized β	IT_4 Shopping Standardized β
RMTscr_1 Tendency to obey a person silently	**.124 (*Pr.* = .015)**	.091 (*Pr.* = .079)	.031 (*Pr.* = .552)	.060 (*Pr.* = .900)
RMTscr_2 Sensation seeking tendency	.061 (*Pr.* = . 230)	**.128 (*Pr.* = .015)**	**.138 (*Pr.* = .008)**	.067 (*Pr.* = .252)
RMTscr_3 Anxiety over relationship	.060 (*Pr.* = . 212)	-.080 (*Pr.* = .106)	.045 (*Pr.* = .359)	-.119 (*Pr.* = .203)
RMTscr_4 Desire to surpass regulations	.056 (*Pr.* = . 273)	.025 (*Pr.* = .630)	**.103 (*Pr.* = .049)**	**.020 (*Pr.* = .016)**
RMTscr_5 Avoidance of physical pain	**.193 (*Pr.* = .000)**	.026 (*Pr.* = .592)	.075 (*Pr.* = .124)	.090 (*Pr.* = .703)
R^2	**.074 (*Pr.* = .000)**	**.024 (*Pr.* = .007)**	**.041 (*Pr.* = .000)**	**.028 (*Pr.* = .027)**

b) Multiple regression analysis using problematic Internet experiences as the dependent variables and level of familiarity with Internet terminology as the independent variables:

We wished to determine to what degree it was possible to predict the four kinds of problematic Internet experiences by focusing on specific Internet jargon. It was suggested that the use of such terms significantly predicted an “addiction tendency (IT_1)” as well as problematic experiences related to “communication (IT_3).”

It was found that the standardized R2 = .104 (Pr. = .000) for “addiction tendency” was significant. Of the items related to the evaluation of familiarity with Internet terminology entered for “addiction tendency,” “the jargon…wwww.” (literally, “laugh, laugh, laugh, laugh”), showing rather scornful laughter about the object in question, was found to be an especially significant independent factor (β = .197, Pr .= .000). Moreover, for the problematic experiences in the category “communication IT_3,” a weak but significant regression using the standardized R2 = .034 (Pr. = .005) was observed, and especially in terms of terminology use, it was suggested that the word “godly,” jargon which is frequently used to praise the interlocutor or others exaggeratedly, could explain problematic communication experiences (β = .144, Pr. = .029) (Table 5).

Table 5 Multiple regression analysis using the total scores for various factors related to problematic Internet experiences (IT_1–IT_4) as the dependent variables and the level of familiarity with Internet terminology as the independent variables (using the forced entry method)

	IT_1 Addiction tendency Standardized β	IT_2 Illegal experience Standardized β	IT_3 Communication Standardized β	IT_4 Shopping Standardized β
W01A Godly	.076 (*Pr.* = .231)	—	**.144 (*Pr.* = .029)**	—
W02A kwsk (tell me more)	.090 (*Pr.* = .124)	—	.088 (*Pr.* = .150)	—
W03A Genius	-.008 (*Pr.* = .892)	—	-.003 (*Pr.* = .965)	—
W04A ○○www (scornful laughter)	**.197 (*Pr.* = .000)**	—	.075 (*Pr.* = .185)	—
W05A meshiuma (literally, "my meal will be delicious," indicating the sentiment of schadenfreude)	.021 (*Pr.* = .753)	—	.065 (*Pr.* = .348)	—
W06A Sore daretoku? (Who will benefit from this?)	-.062 (*Pr.* = .304)	—	-.031 (*Pr.* = .625)	—
W07A orz (disappointed)	.108 (*Pr.* = .055)	—	-.009 (*Pr.* = .883)	—
W08A pgr (sentiment of derision)	-.037 (*Pr.* = .523)	—	.019 (*Pr.* = .754)	—
W11A 888888 (appraisal)	.093 (*Pr.* = .088)	—	-.038 (*Pr.* = .499)	—
R^2	**.104 (*Pr.* = .000)**	n.s.	**.034 (*Pr.* = .005)**	n.s.

c) Single regression analysis using problematic Internet experiences as the dependent variables and the MSSST-A and Internet terminology comprehension scores as the independent variables:

Here, we conducted a single regression analysis to determine to what degree the scores on the normal comprehension for sarcasm, metaphor and Internet terminology could explain the four problematic Internet experiences. It was found that factors that indicated a comprehension of sarcasm, metaphor and jargon significantly predicted "addiction dependency to the Internet, IT_1," "shopping, IT_4," and "communication, IT_3."

It was found that the standardized R2 = .118 (Pr. = .012) for "addiction tendency (IT_1)" was significant. For "addiction tendency," the values of β = -.112 and Pr. = .017 were obtained. The standardized R2 = .013 (Pr. = .017) for "shopping (IT_4)" was significant with the values of β = -.112 and Pr. = .017. For problematic experiences corresponding to "communication (IT_3)," the standardized R2 was not found to be a significant regression coefficient, but the reference value of β = .097 and Pr. = .040 was suggested (Table 6).

Table 6 Single regression analysis using problematic Internet experiences as dependent variables and the MSST-A and Internet terminology comprehension scores as the independent variables (using the forced entry method)

	IT_1 Addiction tendency Standardized β	IT_2 Illegal experience Standardized β	IT_3 Communication Standardized β	IT_4 Shopping Standardized β
Scores for MSST-A and terminology comprehension	**.118 (*Pr.* = .012)**	—	**.097 (*Pr.* = .040)**	**-.112 (*Pr.* = .017)**
R^2	**.012 (*Pr.* = .012)**	n.s.	n.s.	**.013 (*Pr.* = .017)**

4.8 Factors explaining the relationship between private Internet connection time and problematic Internet experiences

A single regression analysis using Internet connection time as the independent variable did not significantly predict any of the problematic Internet experience categories. A simple correlation was calculated between the personal attributes RMTscr, which had been entered so far as independent variables, and the comprehension scores for sarcasm, metaphor and Internet terminology. Internet connection time was found to have a significant correlation with "addiction tendency (IT_1)" (r = .208, Pr. = .000) and the "tendency to obey a person silently (according to an influence from the external environment)" (r = .109, Pr. = .024), among other problematic Internet experience categories.

5 Discussion and Conclusion

At present, the Internet or "cyber space" is an information tool widely used in the social life of children, adolescents, and adults. Few people would deny that a virtual world "exists" in parallel with the real world. Those who have access to the network environment and who have accessibility recognize themselves as "residents" of both the real world and the cyber world. Studies on the developmental processes children go through to learn communication with others have so far assumed a simultaneous interaction with others in a face-to-face manner. Our survey data demonstrated that in the study of communication and cognitive development in such fields as psychology, it is necessary to pay attention to the role of "cyber space" in children's development.

Our survey data suggested several areas that should be included in information literacy education.

1) It has been proven that problematic Internet experiences can be roughly classified into addictive Internet access, tendency to engage in unlawful and illegal behavior, communication errors and misunderstanding, and problems with commercial transactions over the Internet.

2) The survey results suggested that people whose personal attributes make them open to fraud, unscrupulous business, and solicitation risks, have a tendency for collective behavior selection, which includes suggestibility and rising anxiety. Further, a tendency toward anxiety over relationships is because of a fear of hurting others' feelings. The desire for sensation-seeking, however, indicates an egocentric intention to ignore existing regulations, and a tendency to avoid physical activities in the real world.

3) For the comprehension of sarcasm, metaphor, and Internet terminology, we had difficulty developing questions and exercises that students would face in their actual daily life as they might be differently understood and interpreted (including within a communication context on the Internet). However, several questions allowed us to observe deviated samples with a response rate that could be used for a statistical test. It was confirmed that it is possible to predict problematic Internet experiences using sample groups that suggest differences in the understanding of such metaphor, sarcasm and Internet terminology. The "addiction tendency to the Internet, IT_1" and "shopping, IT_4" which were related to shopping experiences over the Internet suggested that scores on the comprehension of sarcasm, metaphor, and Internet terminology had weak but significant explanatory power.

4) Today, "Internet addiction" is an issue that has been recognized as posing a threat to adolescent psychological and social health. This is an addiction category that has received increasing

attention over the last ten years in the fields of clinical and educational psychology and psychiatry. Based on the data obtained from this survey, the profile of "Internet addiction" as a problematic behavior can be described in the following manner. Those who access the Internet addictively, while being also affected by suggestibility and anxiety, also sometimes post messages deriding others, including "…wwww." In general, they tend to choose collective swarm behavior and avoid physical activities (in the real world). It has also been suggested that some tend to select "deviated values" in the comprehension of sarcasm, metaphor, and Internet terminology. Urgent intervention measures are needed if pressure from collective society drives young people to spend more time connected to the Internet than they spend living and interacting with others in the real world.

5) It was also interesting that the "sensation-seeking tendency RMTscr_2," which is considered to be a personal attribute of social risk, was found to be significantly predictive of problems classified as belonging to the category "illegal experiences IT_2," including putting in danger one's personal information or the secrets of their organization. Curiosity and sensation-seeking (fear of boredom) are developmental issues for children and adolescents who extend their sphere of activity from their family to society. Many children and adolescents have the possibility of entering into dangerous worlds out of curiosity or sensation-seeking. This study suggested that this should be taken into consideration in information literacy education programs.

In Japan, high schools and universities deliver courses on information literacy and conduct lectures on Internet use. They teach young people, who are actual users, how to behave prudently, and they present comprehensive examples of potential problems. However, it is generally outside class time that young people access the Internet, which makes it difficult to know whether such information literacy education has any effect on improving their knowledge and self-management capacity.

To further reduce the risks related to Internet use, educational intervention aimed at promoting users' awareness of their personal behavioral characteristics and intentions should be provided. If students were aware of the risks they expose themselves to by going through a self-analysis process prior to Internet use, this may reduce their propensity for risky behavior.

This study was conducted on female students on the basis of self-reflection by the respondents. By performing surveys which target users of various ages, we could examine in more detail the factors related to the problems young people have when using the Internet as well as the personal characteristics of those who could be more likely to experience such problems.

In surveys based on self-report responses, there is a possibility that the problems encountered on the Internet might not be reflected unless the participants recognized these as problems. Further, the presence of samples where the responses were based on what the teacher wanted cannot be excluded due to the survey method used in this study. Nevertheless, differences between individuals as regards the understanding of "sarcasm," "metaphor," and "specific Internet terminology" and in the number of problems encountered were observed. This study has been valuable in that clues to the risk factors for "Internet problems" have been suggested.

In the future, it will become necessary to develop educational tools that could be used to estimate the risk of Internet use and to call attention in a more concrete and useful way to each individual's behavioral attributes by collecting more examples and additional survey data. It is necessary to establish information literacy education programs aimed at helping adolescents become aware of the necessity of protecting themselves and of achieving skills for safe Internet use.

This work was supported by JSPS Grant Number 25330429.

REFERENCES

[1] OECD: THE PROTECTION OF CHILDREN ONLINE: Risks faced by children online and policies to protect themtal Psychology, 14, 19asses of Collective Attention in Twitter. In Proceedings of the 21st International Confere(accessed 2014-12-17)

[2] Ministry of Internal Affairs and Communications Institute for Information and Communications Policy. Announcement of Issuance of FY2013 Internet Literacy Indicator for Students Etc. http://www.soumu.go.jp/main_content/000175589.pdf（accessed 2014-12-17）.

[3] HIRAI Tomohisa. (2012). Why does "Enjyo" happen on the Web? : An Examination based on Japanese Web Culture, Journal of Information and Communication Research, Vol.29 No.4, 61–71.

[4] Thurlow, C., Lengel, L., and Tomic, A. (2004). 004e .mic, 18101ification method of Internet related troubles,MT) as a new

[5] O'Sullivan, P.B. and Flanigan, A.J. (2003). Reconceptualizing 'Flaming' and Other Problematic Messages, New Media & Society, Vol.5 (1): 69–94, SAGE.

[6] Takeichi, Y., Sasahara,K., Suzuki, R., and Arita,T. (2014). Twitter as Social Sensor: Dynamics and Structure in Major Sporting Events, Artificial Life 14, 778–784.

[7] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a Social Network or a News Media? In Proceedings of the 19th International Conference on World Wide Web, 591–600.

[8] Bollen, J., Goncalves, B., Ruan, G., and Mao, H. (2011a). Happiness is Assortative in Online Social Networks. Artificial Life, 17(3):237–251.

[9] Gabowicz, P. A., Ramasco, J. J., Moro, E., Pujol, J. M., and Egu´ıluz, V. M. (2012). Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media.PLoS ONE, 7(1);e29358

[10] Conover, M. D., Gonc¸alves, B., Flammini, A., and Menczer, F. (2012). Partisan Asymmetries in Online Political Activity. EPJ Data Science, 1(1):6.

[11] Romero, D. M., Meeder, B., and Kleinberg, J. (2011). Differencesyrne, & Whiten, A. (Eds.), Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans. Oxford: Clarendoternational Conference on World Wide Web, 695–704.

[12] Weng, L., Flammini, A., Vespignani, A., and Menczer, F. (2012). Competition among memes in a world with limited attention. Scientific Reports, 2:335.

[13] Lehmann, J., Gonc¸alves, B., Ramasco, J. J., and Cattuto, C. (2012). Dynamical Classes of Collective Attention in Twitter. In Proceedings of the 21st International Conference on World Wide Web, 251–260.

[14] Golder, S. A. and Macy, M.W. (2011). Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. Science, 333(6051):1878–1881.

[15] Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. In Proceedings of the 19th International Conference on World Wide Web, 851–860.

[16] Borge-Holthoefer, J., Rivero, A., Garc´ıa, I., Cauh´e, E., Ferrer, A., Ferrer, D., Francos, D., I˜niguez, D., P´erez, M. P., Ruiz, G., Sanz, F., Serrano, F., Vi˜nas, C., Taranc´on, A., and Moreno, Y. (2011). Structural and Dynamical Patterns on Online Social Networks: The Spanish May 15th Movement as a Case Study. PLoS ONE, 6(8):e23883.

[17] Gonz´alez-Bail´on, S., Borge-Holthoefer, J., Rivero, A., and Moreno, Y. (2011). The Dynamics of Protest Recruitment through an Online Network. Scientific Reports, 1:197.

[18] Takahashi, T., Tomioka, R., and Yamanishi, K. (2014). Discovering Emerging Topics in Social Streams via Link-Anomaly Detection. Knowledge and Data Engineering, IEEE Transactions on, 26(1):120–130.

[19] Bollen, J., Mao, H., and Zeng, X. (2011b). Twitter Mood Predicts the Stock Market. Journal of Computational Science, 2(1):1–8.

[20] Sasahara, K., Hirata, Y., Toyoda, M., Kitsuregawa, M., and Aihara, K. (2013). Quantifying Collective Attention from Tweet Stream. PLoS ONE, 8(4):e61823.

[21] Heinrich, HW. 1931. Industrial Accident Prevention. New York: McGraw-Hill.

[22] Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? The Behavioral and Brain Sciences, 1, 515–526.

[23] Premack, D. (1988). 'Does the chimpanzee have a theory of mind?' revisited. In R. Byrne, & Whiten, A. (Eds.), Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans. Oxford: Clarendon Press. 160–179.

[24] Hobson, R.P. (1993). Autism and the development of mind. Hove: Lawrence Erlbaum Associates.

[25] Moore, C. (1996). Theories of mind in infancy. British Journal of Developmental Psychology, 14, 19–40.

[26] Simon Baron-Cohen, Alan M. Leslie, Uta Frith. (1985). Does the autistic child have a “theory of mind” ?, Cognition, Volume 21, Issue 1, 37–46.

[27] Takamiya, Matsui et al (2009). Brain activation associated with theory of mind: An fMRI study. Journal of Human Environmental Studies Vol. 7 (2009) No. 2 P 129–135.

[28] Moriguchi, Y. 2011 Psychosomatic medicine and neuroscience on understanding of selves and other's mind. Journal of the Human Arts and Sciences Vol. 7 (2011) No. 1 10–16.

[29] Adachi, T. (2006). The Study of Situational Recognition of Attention Deficit/Hyperactivity Disorders, Asperger's Disorder and High Functioning Autism with the Metaphor and Sarcasm Scenario Test (MSST), Official Journal of the Japanese Society of Child Neurology 38(3), 177–181, 2006-05-01.

[30] Takeuchi, M., Suzuki, M. (2000). The basic study of the Self Risk Management Test (SRMT) as a new tool for the educational programs informing lifetime risks of swindling victimization, Sanno College Junior Bulletin, Vol.33, 11–24.

[31] Maslow, A.H. (1954). Motivation and Personality, Harper & Row, NY.

[32] Cloninger, C.R. (1987). A systematic method for clinical description and classification of personality varients: a proposal. Arch. Gen. Psychiat., 44, 573–588.

[33] Tashiro, M. (2011). Proposal of classification method of Internet related troubles, The Infosocionomics Society Vol.6, No.1, 101 – 114, 2011-6-18.