International Journal of Computer Applications Technology and Research
Volume 3– Issue 2, 88 - 92, 2014

A New Approach to Segmentation of On-Line Persian Cursive Words

Golestani, M.R
Department of Computer Engineering, Islamic Azad University South Tehran Branch
Tehran, Iran

Khademi, M.
Department of Applied Mathematics, Islamic Azad University South Tehran Branch
Tehran, Iran

Moeini, A.
Division of Algorithms and Computation, Department of Engineering Science,
College of Engineering,
University of Tehran
Tehran, Iran

Abstract: Segmentation approaches, as processes that divide word into smaller parts which contain one letter at most, have important effect on cursive word recognition. While online cursive word recognition became applied technology in Latin and Chinese languages, complex structural features in Arabic-based script made it an important field of study in Persian and Arabic languages. In this paper, by introducing of Standard Persian Handwriting, we proposed a novel approach to segmentation online Persian cursive script based on width of letter's body in Persian language. Results are shown 99.86% accuracy in detection of expected segmentation points, while recognized extra points reduced 93.73% compared to our previous methods.

Keywords: Segmentation, Persian cursive handwriting, Letters width, Words feature, Online recognition

1. INTRODUCTION

In recent decade, by growth in usage of Pen-Based devices such as Smart phones, Tablets and etc., online handwriting recognition get lots of attentions. Online cursive word recognition studies gain more success in Latin and Chinese languages due to frequent researches [1, 2, 3], while complex structural features in Arabic-based script made online Persian cursive recognition an open field of study.

In Arabic and Persian, words are cursive by nature and every letter could have different shapes based on position of the letter in a word. According to these features and some other like different position of diacritic marks on letters with common body, recognition of Persian and Arabic cursive words are difficult. In recent years, researchers tried to reduce this complexity [4-12].

For reducing complexity, we could break up whole word recognition to smaller parts. Segmentation is a way to divide cursive word into smaller parts which contain one letter at most. By usage of segmentation, one could recognize whole word by combination of small parts recognition result. So segmentation method has important effect on whole word recognition. Recently, different efforts made to propose segmentation method in online Persian and Arabic cursive word recognition [4-11, 13].

In this paper, by introducing of Standard Persian Handwriting, a new approach has been proposed for detecting the segmentation points. The rest of the paper is organized as follows: In next section, we will have a review on researches have been made upon the subject. In Section 3, we introduce Standard Persian Handwriting. In Section 4, we introduce our segmentation method. Section 5, illustrates the experimental results and finally, the paper will be concluded in section 6.

2. LITERATURE SURVEY

Due to effect of segmentation on whole cursive word recognition in Persian and Arabic language, various segmentation methods were presented to reduce recognition complexity by researchers. A list of Persian language features was introduced and used to find segmentation points [6]. By usage of writing direction, points were considered that follow a specific pattern in their direction before and after themselves, as a segmentation point. Series of directions is another feature that was used [6, 7]. The last repeating points in a series of adjutant points, first and last points of each strokes of a word, point's first and second derivatives of right and left, are other features that were used to detect segmentation points in [6].

In [8], according to curve structures in Persian language, curves and their features was detected in input points, then segmentation points defined based on them. Samimi et. all [7] convert input points to their proposed patterns. They defined 7 basic shape including Semicircle in 4 direction, horizontal line, vertical line and oblique line. After conversion of input points to related pattern, segmentation points were detected based on shapes.

Khaled et. all [10] studied on Arabic cursive word. They first defined joint line by angle between line of adjacent points and the horizontal axis and only consider semi-horizontal lines moving from right to left. Then checked above and below of the joints and keep joints that have not any point in above and below. Finally, by integrating these joint lines, middle points were considered as segmentation points.

Two specific features in Persian words were defined in our previous proposed method [13]. In joining of two letters, joining position always be placed over the right-to-left writing direction. Also there is an obvious difference in gradient of ending points of the preceding and beginning points of the succeeding letter. With usage of these features, they first found points with specific gradient on right-to-left direction, then track this points to detect beginning point of next letter and consider it as segmentation point. While segmentation points were detected in this method, some other extra points were detected too.

3. STANDARD HANDWRITING

Persian alphabet contains 32 letters, which has 4 more letters than Arabic. Letters in Persian and Arabic languages could have different shapes based on position of the letter in a word.

Every letter could have utmost four different shapes for isolated, initial, middle and final modes. All different shapes of all Persian letters illustrated in [12].

In addition, various and different writing styles in Persian handwriting, made Persian handwriting recognition complicated. In some samples people write letters totally different from letter main structure or in some other they deform a letter and turn it to some other letters. For example they write letter "د" same as letter "ر ". In some other samples letters were written inside each other and have overlap.

To prevent this undesirable complexity, we introduce a Standard Persian Handwriting that will be applied on handwriting samples. A standard Persian Handwriting must include 3 below condition:

- Letter shape must be written same as normal Persian letter shape. For example letter "د" must not write same as letter "ر".

- Joining letters must have not overlap.

- Words must be written on a straight horizontal line. It means that each part of word did not write in different position. Parts of word must be written along each other same as normal Persian language.

So, if someone consider normal Persian language and do not create a new shape for letters, handwriting will be Standard Persian Handwriting. In this research, all writers followed Standard model and proposed method is based on Standard Persian Handwriting samples.

4. PROPOSED ALGORITHM

In this Section, by introducing a special feature in Persian language, we present a new method for detection of segmentation points in Persian cursive words. In the following, we first depict the feature, and then present our algorithm.

4.1 The Feature

Width of letters body in Persian language has difference in different parts of the letter. If we just consider beginning part of letter, usually have bigger width than situation that we just consider end part of letter. Figure 1 shows different widths in various parts of a word. In joining of two letters, joining position has lowest width as shown in Figure 1. This feature was confirmed for all Persian letters joining. Joining position is best candidate for segmentation points because segmentation process goal is to divide words to letters or basic shapes.

According to this feature, one could detect all positions that letters joint together in a word. Although some other parts of words could include same feature, too.

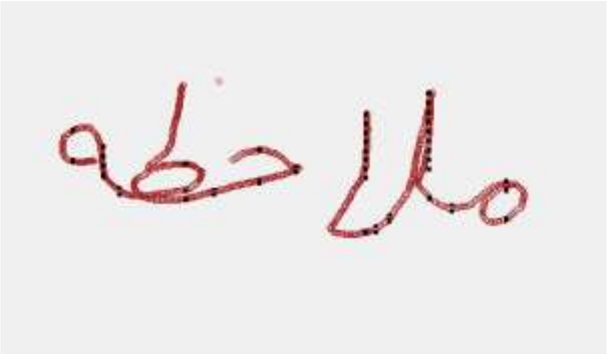


Figure. 1 Different width in various parts of word "ملاحظه"

4.2 Segmentation Algorithm

In on-line handwriting, when somebody writes a word on the screen, simultaneously, the handwriting is stored as series of input points, including X and Y coordinates. As we put the pen tip down until picking it up, its movement is a continuous curve, containing series of points, which we define it as a stroke. Persian words could contain one or more stroke. For identifying of segmentation points a specific word, we need to detect all the segmentation points for each stroke [13].

Our new suggested method include 2 phase. In first phase, segmentation points are detected according to described feature. Then, in second phase, some extra points are removed by post-processes.

In first phase, following steps execute for all strokes in a word:

- A grid, including rows and columns, is drawn around the word. each small square part of grid called a Cell. Cells dimension defined by a threshold.

- If there is an input point in a cell position, that cell will be called Active cell. For all points of stroke, related cell turned to Active cell.

- Columns with only one active cell are detected. Consecutive columns with one active cell considered as integrated set. Each set, whether it is integrated set or one column, is considered as a candidate line.

- If two candidate lines are separated by one column, number of active cell on that column must be check. If number of active cell, width, is 2 then two candidate line will be joined together and will be considered as one candidate line.

- Most left column of each candidate line, consider as segmentation area. Last point of stroke that is placed in this column, will be detected as segmentation point.

By execution of these steps, all points that have described feature will be detected. To avoiding input error, in step 4, we check width of column that separated two candidate lines. If it was small, we reject it and join two candidate lines together.

In second phase, some specific extra points are removed by post-processes. Following post-processes execute, after first phase completion:

- By usage of horizontal projection concept, word baseline is detected. So by using of drawn gird, the raw that has most active cells will considered as baseline. In several researches Horizontal Projection was used to detect Persian words baseline [14].

- Considering last detected segmentation point in each stroke. If the detected point is placed below of baseline, 4 time more than threshold, while last point of stroke is placed around baseline, the detected segmentation point is considered as extra point in end of letters such as "ل" ,"ن" and etc. and will be removed.

- Considering last detected segmentation point in each stroke. If the detected point is placed around baseline, while last point of stroke is placed around same column and below of baseline, 5 time more than threshold, the detected

Table 2. Result of detected extra points by proposed method

Words	Number of words recognized without extra point	Number of words recognized with 1 extra point	Number of words recognized with 2 extra point	Number of words recognized with 3 extra point	Number of words recognized with 4 extra point
هرچند	48	2	0	0	0
تلافی	46	4	0	0	0
عاشق	42	8	0	0	0
بصیرت	49	1	0	0	0
کرج	35	13	2	0	0
مریض ترسناک	12	25	8	5	0
فلسطین	42	8	0	0	0
ملاحظه	50	0	0	0	0
گهگاه	15	17	12	5	1
تفریط	50	0	0	0	0
20 other words	834	159	7	0	0
Sum	1223	237	29	10	1

segmentation point is considered as extra point in end of the letter "م" and will be removed.

- Considering last detected segmentation point in each stroke. If the detected point is placed right side of last point of stroke, less than quadruple threshold, while end point of stroke is placed top of detected point, less than sextuple threshold, the detected segmentation point is considered as extra point in end of letters such as "ب" ,"ت" and etc. and will be removed.

- Considering last detected segmentation point in each stroke. If the detected point is placed below of baseline, 3 time more than threshold, and is one of four last stroke points, probably it is an extra point on the letter "ر". So previous points are tracked to find a point around baseline, the point is called candidate point. If there is not another segmentation point around candidate point, then detected point is considered as extra point in end of "ر" and will be removed and candidate point is considered as last segmentation point of stroke, otherwise just detected point is considered as extra point and will be removed.

- In last post-process, detected segmentation points that are placed in last or first column of a stroke, are considered as extra points and will be removed. In first phase, these points were placed in an isolate cell, so they were detected as segmentation points by mistake.

By execution of these post-processes, most of specific extra points will be removed and only correct segmentation points will be remained. Figure 2 shows detected segmentation points of the word "ملاحظه" after execution of proposed method.

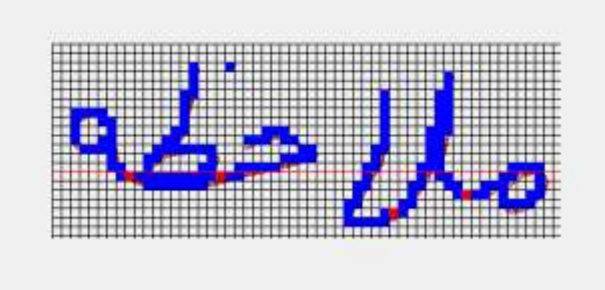


Figure. 2 Detected segmentation points of the word "ملاحظه" by proposed method

5. EXPERIMENTAL RESULTS

To evaluate the efficiency of the proposed method, we developed a program to take people handwriting by a digital pen and detect segmentation points of them by our proposed method. We used Microsoft Visual Studio 2010 programming Environment and C# programming language to develop it.

In selection of the words, we picked out 30 words that include all structures of Persian letters. We ask 50 different people to write the words by considering Standard Persian Handwriting that we described before. Table 1 presents the results of the proposed method.

Table 1. Result of detected segmentation points by proposed method

Words	Number of expected point In 50 sample	Number of detected points	Number of undetected points
هرچند	150	149	1
تلافی	150	149	1
عاشق	200	199	1
بصیرت	200	199	1
کرج	50	49	1
مریض ترسناک	400	398	2
فلسطین	350	350	0
ملاحظه	200	200	0
گهگاه	150	150	0
تفریط	150	150	0
20 other words	3150	3150	0
Sum	5150	5143	7

The experimental results show that in 5 words, only in one handwriting sample from 50 samples, one segmentation point does not detect. And also for the word "مریض ترسناک", only in two handwriting sample from 50 samples, one segmentation point does not detect. All other expected segmentation points recognized correctly by our proposed method.

Table 3. Result of detected segmentation points by our previous method

Words	Number of expected point In 50 sample	Number of detected points	Number of undetected points
هرچند	150	150	0
تلافی	150	149	1
عاشق	200	200	0
بصیرت	200	200	0
کرج	50	50	0
مریض ترسناک	400	400	0
فلسطین	350	349	1
ملاحظه	200	200	0
گهگاه	150	149	1
تفریط	150	150	0
20 other words	3150	3148	2
Sum	5150	5145	5

In addition, according to the experimental results, our proposed method could detect some extra points too. Based on the obtained result of detected extra points that is shown in Table 2, 81.53% of words recognize without any extra point and 15.8% of words recognize with one extra point and 2.67% of words recognized with 2, 3 or 4 extra points.

Also, we used handwriting samples to evaluate our previous method [13]. To improve that method, we first filtered input points and only kept adjacent points that have interspace longer than a threshold. Then, we applied our previous method [13] on filtered points. The results of our previous method for detected segmentation points and detected extra points were shown in Table 3 and Table 4.

Table 4. Result of detected extra points by our previous method

Words	Number of words recognized without extra point	Min Number of detected extra point	Max Number of detected extra point
هرچند	0	1	9
تلافی	0	1	8
عاشق	0	1	8
بصیرت	0	1	9
کرج	0	1	10
مریض ترسناک	0	1	10
فلسطین	0	1	10
ملاحظه	0	1	6
گهگاه	0	1	19
تفریط	10	1	8
20 other words	7	1	18
Sum	17		

The obtained result showed that while previous method could recognize 99.9% of segmentation points, it recognize many extra points too. Only 1.13% of words recognized without any extra points and most of words were detected with more than 3 extra points.

Table 5 shows total number of detected extra points for proposed method and our previous method. The result showed that proposed method reduced detected extra points 93.73% compared to previous [13] method.

Table 5. Total number of detected extra points by proposed method and previous method

	Total number of detected extra points
Proposed method	329
Previous [13] method	5249

6. CONCLUSIONS AND FUTURE WORKS

In this paper, due to different width in various parts of Persian letters, we proposed a new approach to detecting of segmentation points in on-line Persian cursive script words. At first, we reviewed some related works and our previous method. Then Standard Persian Handwriting was introduced. Later, specific feature of Persian language was described as well as our proposed method. The experimental results from evaluation of our algorithm have shown that 99.86% of expected segmentation points, means the last point of each letter in a joining or last point of basic shapes, are detected. Proposed method detected 81.53% of words without extra points. Comparison has indicated that proposed method reduced detected extra points 93.73% compared to our previous method.

In future works, to improve proposed method, according to detected extra points in diacritic marks of words same as slant and etc., we could separate main body of word and diacritic marks and reduced number of extra points. For future work, according to 99.86% accuracy in detection of segmentation points, we could try to recognize each detected segment by isolate Persian character recognition approaches or HMM and finally complete on-line Persian cursive word recognition.

7. REFERENCES

[1] Yuan, A., Bai, G., Yang, P., Guo, Y., & Zhao, X. 2012. Handwritten English Word Recognition based on Convolutional Neural Networks. In Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on (pp. 207-212). IEEE.

[2] Liu, C. L., Yin, F., Wang, D. H., & Wang, Q. F. 2012. Online and offline handwritten Chinese character recognition: benchmarking on new databases. Pattern Recognition.

[3] Dai, R., Liu, C., and Xiao, B. 2007. Chinese character recognition: history, status and prospects. Frontiers of Computer Science in China, 1(2), 126-136.

[4] Biadsy, F., El-Sana, J., and Habash, N. 2006. Online arabic handwriting recognition using hidden markov models. In Tenth International Workshop on Frontiers in Handwriting Recognition.

[5] Halavati, R., Jamzad, M., and Soleymani, M. 2005. A novel approach to persian online hand writing recognition. In Proceedings of the 4th World Enformatika Conference (WEC 05). Vol. 6, pp. 232-236.

[6] Pirnia, Sh., Khademi, M., Nikookar, A. and Bani, Z. 2010. A Feature-Based Approach to Segmentation of Persian Online Cursive Script. The 2010 International Conference on Computer and Software Modeling (ICCSM).

[7] Daryoush, K. S., Khademi, M., Nikookar, A., & Farahani, A. 2012. Segmentation of Persian Cursive Words Using Basic Shapes. Journal of Engineering Research and Applications (IJERA).

[8] Izadi, S., Haji, M., and Suen, C. Y. 2008. A new segmentation algorithm for online handwritten word recognition in Persian script. In Proc. Eleventh International Conf. Frontiers in Handwriting Recognition (CFHR 2008) . pp. 598-603.

[9] Harouni, M., Mohamad, D., & Rasouli, A. 2010. Deductive method for recognition of on-line handwritten Persian/Arabic characters. In Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on IEEE. Vol. 5. 791-795.

[10] Daifallah, K., Zarka, N., & Jamous, H. 2009. Recognition-based segmentation algorithm for on-line arabic handwriting. In Document Analysis and Recognition. 10th International Conference on. IEEE. 886-890.

[11] Maliki, M., Jassim, S., Al-Jawad, N., & Sellahewa, H. 2012. Arabic handwritten: pre-processing and segmentation. In Proc. of SPIE. Vol. 8406.

[12] Harouni, M., Mohamad, D., Rahim, M. S. M., and Halawani, S. M. 2012. Finding Critical Points of Handwritten Persian/Arabic Character. IJMLC.

[13]] Khademi, M., Golestani, M.R., Nikookar, A., and Farahani, A. 2013. A New Method for Detecting Segmentation Points in Persian Cursive Words. Journal of Basic and Applied Scientific Research (JBASR),Special Issue (1).

[14] Nagabhushan, P., and Alaei, A. 2010. Tracing and straightening the baseline in handwritten persian/arabic text-line: A new approach based on painting-technique. The Proceeding of Intl Journal on Computer Science and Engineering. 907-916.

Impact of Bio-inspired metaheuristics in the data clustering problem

K. Sumangala,
Assistant Professor, Research Scholar,
Department of Computer Science,
Kongunadu Arts & Science College,
Coimbatore - 641 030, India
.

K. Papitha
Department of Computer Science,
Prince Shri Venkateswara Arts & Science
College,Chennai
India

Abstract: The goal of data mining is to extract the knowledge from data. It is also a form of knowledge discovery essential for solving problem in a specific domain. This paper presents a novel approach to data clustering and classification problem. Clustering analysis is distribution of data into groups of similar objects and Classification focuses the data on the class boundaries. This research explores three different bio-inspired metaheuristic algorithms in the clustering problem: Ant Colony Optimization (ACO), Genetic Algorithms (GAs) and Artificial Immune Systems (AIS). Data mining approaches are applied in the field of medical diagnosis recently. The major class of problem in medical science involves diagnosis of disease based upon various tests. The computerized diagnostic tools are helpful to predict the diagnosis accurately. Breast cancer is one the most dangerous cancer type in the world. Early detection can save a life and increase survivability of the patients. This of research work analysed the performance of GA, ACO and AIS with ID3 for solving data clustering and classification problem in an experiment with Breast Cancer Dataset data of UCI repository. An efficient ID3 Decision tree based classification techniques are used to measure the performance of the system with GA, ACO and AIS system. Proposed AIS system produces the best classification result than the ACO and GA based decision tree ID3 classifiers. Instead of K-means clustering, this research work combines the simplicity of K-means algorithm with the robustness of AGA-Miner. This proposed approach has potential applications in hospital for decision-making and analyze/ research such as predictive medicine.

Keywords: Clustering problem; genetic algorithms; ant colony optimization; artificial immune systems, ID3 classification techniques, UCI data repository.

1. INTRODUCTION

The goal of data mining is to extract the knowledge from data. It is also a form of knowledge discovery essential for solving problem in a specific domain. This paper presents a novel approach to data clustering and classification problem. Clustering analysis is distribution of data into groups of similar objects and Classification focuses the data on the class boundaries. This research explores three different bio-inspired metaheuristic algorithms in the clustering problem: Ant Colony Optimization (ACO), Genetic Algorithms (GAs) and Artificial Immune Systems (AIS). Data mining approaches are applied in the field of medical diagnosis recently. The major class of problem in medical science involves diagnosis of disease based upon various tests. The computerized diagnostic tools are helpful to predict the diagnosis accurately. Breast cancer is one the most dangerous cancer type in the world. Early detection can save a life and increase survivability of the patients. This of research work analysed the performance of GA, ACO and AIS with ID3 for solving data clustering and classification problem in an experiment with Breast Cancer Dataset data of UCI repository. Classification Trees are methodologies to classify data into discrete ones using the tree-structured algorithms. It uses information gain to select best attribute for splitting. An efficient ID3 Decision tree based classification techniques are used to measure the performance of the system with GA, ACO and AIS system. Proposed AIS system produces the best classification result than the ACO and GA based decision tree ID3 classifiers. Instead of K-means clustering, this research work combines the simplicity of K-means algorithm with the robustness of AGA-Miner. This proposed approach has potential applications in hospital for decision-making and analyze/ research such as predictive medicine. In this study we have developed AGA-Miner that selects the best cluster centroid value than the existing clustering methods.

The rest of this paper is organized as follows. Section 2 presents the data clustering problem, standard K-means algorithm and related Data mining techniques. Section 3 addresses the bio-inspired metaheuristics: ACO, GA, AIS and ID3. Section 4 explains empirical studies performed using these metaheuristics on data clustering and Section 5 compares the different approaches of AGA-Miner. Finally, concluding remarks are given in Section 6.

2. CLUSTERING ALGORITHM

The process of grouping a set of abstract objects into classes of similar objects called clustering [1]. The clustering problem can be described as follows:

$$J(W,C) = \sum_i^N \sum_j^K w_{ij} \,||x_i - c_j||^2 \qquad (2.1)$$

$$\sum_j^k w_{ij} = 1 \qquad (2.2)$$

$$C_j = \frac{1}{N_j} \sum_{\pi \in Cj} x_i \qquad (2.3)$$

Where, K-> number of cluster, n-> number of objects , m-> number of attribute, Cj->center of jthcluster, x_i->location of ith object.

Euclidean distance:
The distance between the data vector and centroid C is calculated by:

$$\sqrt{\sum_{i=1}^{n}(x_i - c_i)^2} \qquad (2.4)$$

2.1. K-Means Algorithm:

One of the most widely used algorithms is k-means clustering. It partitions the objects into clusters by minimizing the sum of the squared distances between the objects and the centroid of the clusters. The k-means clustering is simple but it has high time complexity, so it is not suitable for large data set. One of the most popular clustering techniques is the k-means clustering algorithm.

Starting from a random partitioning, the algorithm repeatedly (i) computes the current cluster centers (i.e. the average vector of each cluster in data space) and (ii) reassigns each data item to the cluster whose center is closest to it.

The algorithm for the standard k- means clustering is given as follows [2]:

a. Choose a number of clusters k

b. Initialize cluster centers µ1,… µk

i. Could pick k data points and set cluster centers to these points

ii. Or could randomly assign points to clusters and take means of clusters.

c. For each data point, compute the cluster center it is closest to (using some distance measure) and assign the data point to this cluster.

d. Re-compute cluster centers (mean of data points in cluster)

e. Stop when there are no new re-assignments.

Advantage:

- Simple and widely used clustering algorithm.

Disadvantage:

- Time complexity (when large dataset are to be clustered)
- Need to estimate the number of cluster in advance.

2.2. Data Mining Techniques

This research work presents a novel approach to data clustering and classification problem [2]. The clustering is to discover the data distribution. The proposed system is able to cluster real value data efficiently and correctly, dynamically estimating number of cluster. This research work analyzed the performance of ACO, GA and AIS metaheuristics algorithms. In classification problem discrimination among classes is based on the decision tree ID3 classifier.

- Decision Tree (ID3)
- Genetic Algorithm (GA)
- Ant Colony Optimization (ACO)
- Artificial Immune System (AIS)

Decision Tree (ID3):

- Iterative Dichotomiser 3 (ID3) is algorithm for building decision tree.
- It uses information gain to select best attribute for splitting.

Genetic Algorithm (GA):

- **GA** is a technique for solving the clustering problem.
- **GKA** combines K-Means with GA to find globally partition for dataset into number of cluster.
- It uses process such as population, crossover, mutation.
- ID3 algorithm applied to classify the cancer dataset result from the cluster result. ID3 selects the test attribute based on Information Gain.

Ant Colony Optimization (ACO):

- It combines K-means with ACO to improve the k-means in two steps:
    - To avoid local optima
    - ACO applied to refine the cluster to improve quality.
- Ants are used to cluster the data points.
- Only one ant is used to refine the cluster.
- Whenever it crosses a cluster, it will pick an item from the cluster and drop it into another cluster while moving with the help of pickup and drop up probabilities.

Artificial Immune System (AIS):

- AIS is try to imitate real immune system. Most AIS use only the main ideas of real immune systems, namely clonal and negative selection which deal with the evolution of B-cells and T-cells, respectively.
- The proposed method belongs to the method derived from immune system paradigm, called ClonalG Selection.
- It resembles the original K-Means algorithm, but it get rid of its main drawback
    - It is able to estimate the proper number of cluster & avoids getting stuck in inappropriate areas.

- Comparing to other immune algorithms for data clustering, its computational cost is decreased by producing a limited number of clones and proper suppression mechanism.

3. METAHEURISTICS GA, ACO, AIS AND ID3 BASED METHODS

Heuristics refers to experience-based techniques for problem solving, learning, and discovery [1]. In computer science, metaheuristicis a computational method that optimizes a problem by iteratively trying to improve a candidate solution. Metaheuristics allows us to find the best solution over a discrete search-space.

3.1 Classification Techniques (ID3):

The decision tree is constructed with each non-terminal node representing the selected attribute on which data was split and terminal nodes representing the class label of the final subset of its branch. ID3 is an algorithm for building decision tree. It uses information gain to select best attribute for splitting. Classification Trees are methodologies to classify data into discrete ones using the tree-structured algorithms [17]. The main purpose of decision tree is to expose the structural information contained in the data. If the target variable (also called as response variable or class variable) is nominal/categorical variable is called "classification tree" and if continuous, the tree is called "regression tree". ID3 is a recursive process used to construct decision tree from data.

Building decision tree:

1. Calculate the entropy for every attribute using the data set S.
2. Split the set S into subsets using the attribute for which entropy is minimum. Make a decision tree node containing that attribute.
3. Recurse on subsets using remaining attributes. Stop splitting when all examples at a node have the same labels.

Entropy:

- Entropy is a quantitative measurement of the homogeneity of a set of examples.
- It tells us how well an attribute separate the training examples according to their target classification.
- Given a set S with only two class case (malignant & benign)

$$\text{Entropy}(S) = -P_m \log_2 P_m - P_b \log_2 P_b \qquad \textbf{(3.1.1)}$$

Where P_m = proportion of malignant examples
P_b = proportion of benign examples

If entropy(S) = 0, all members in S belongs to one class.
If entropy(S) = 1(max value), members are Split equally between two classes.

In general, if an attribute takes more than two values

$$\text{Entropy}(S) = \sum_{i=1}^{n} -p_i \log(p_i) \qquad (3.1.2)$$

P**seudo code of ID3 :**

```
ID3 ( Learning Sets S, Attributes Sets A, Attributes values V)

Begin
        Load learning sets first, create decision tree root node
'rootNode', add learning set S into root node as its subset.
For root Node, we compute Entropy (rootNode.subset) first

        If Entropy(rootNode.subset)==0, then
                rootNode.subset consists of records all with
                the same value for the categorical attribute,
                return a leaf node with decisionattribute:
                attribute value;

        If Entropy(rootNode.subset)!=0, then
                compute information gain for each attribute
                left(have not been used in splitting), find
                attribute A with Maximum(Gain(S,A)).

                Create child nodes of this rootNode and add
                to rootNode in the decision tree.

        For each child of the rootNode, applyID3(S,A,V)
        recursively until reach node that has entropy=0 or reach
        leaf node.
End ID3.
```

- Looking for which attribute creates the most homogeneous branches

$$G\boldsymbol{ain}(\boldsymbol{S},\boldsymbol{A}) = \boldsymbol{Entropy}(\boldsymbol{S}) - \sum_{\boldsymbol{v}\in \boldsymbol{Value}(\boldsymbol{A})} \left|\frac{\boldsymbol{S_v}}{\boldsymbol{S}}\right| \boldsymbol{Entropy}(\boldsymbol{S_v}) \quad (3.1.3)$$

Where, A is an attribute of S, Value(A) is the set of possible value of A, v is a particular value in Value(A), S_v is a subset of S having of v's on value(A)

3.2. GA with clustering & classification:

- **GA** is a technique for solving the clustering problem [8].
- **GKA** combines K-Means with GA to find globally partition for dataset into number of cluster [1]. The purpose of GKA (Genetic K-Means Algorithm) is to minimize intra-cluster variance.
- From the initial population, the basic operators (selection, crossover, mutation) evolve the population generation to generation. These operators has been used to populate the data

points which helps to find the best fitness solution.

- **Operations:**
    - **Population:** Initially, populate the Breast Cancer dataset attribute value as 0's and 1's.
    - **Selection: "Select the best, discard the rest".** Select the present attribute data & compute the fitness solution. The selection operations, selects the best fitness solution and stores into global.
    - **Crossover & Mutation:**
        - **Crossover:** Cross over operator combines parts of two part solutions to create new solution.
        - **Mutation:** Mutation operator modifies randomly the solution created by crossover.

Euclidean distance:

The distance between the data vector x and centroid c is calculated by,

$$\sqrt{\sum_{i=1}^{n}(x_i - c_i)^2} \qquad (3.2.1)$$

Pseudo code of GA-Clustering Algorithm

```
Begin
    1.  t=0
    2.  Initialize population P(t)
    3.  Compute fitness P(t)
    4.  t=t+1
    5.  If termination criterion is achieved go to step 10
    6.  Select P(t) from P(t-1)
    7.  Crossover P(t)
    8.  Mutate P(t)
    9.  Go to step 3
    10. Output best and stop

End
```

- ID3 algorithm applied to classify the cancer dataset result from the cluster result.

- ID3 selects the test attribute based on Information gain. IG measures the change of uncertainty level after classification from an attribute.

3.3 ACO with clustering & classification:

The proposed system combines K-Means with Ant Colony Optimization to improve K-Means in two steps [7]:

1. To avoid local optima.
2. ACO applied to refine the cluster to improve quality.

Pseudo code of ACO Algorithm

```
ACO Algorithm:

    1. Choose number of cluster k
    2. Initialize cluster center $\mu_1, \mu_2 \ldots\ldots \mu_n$.
    3. For each data points , compute the cluster center ,it is
       closet to and assign the data point to this cluster.
    4. Re-compute cluster center.
    5. Stop when no new reassignments.
    6. Ant based refinement:
        1. Input the cluster from improved K-means.
        2. For i=1 to N do
            1. Let the ant go for random walk to pick an item.
            2. Calculate pick up & drop up probability.
            3. Decide to drop the item.
            4. Re-calculate the entropy value to check whether the
               quality is improving.
        3. Repeat.
```

The ants are used to cluster the data points. Here, only one ant is used to refine the cluster. Whenever it crosses a cluster, it will pick an item from the cluster and drop it into another cluster while moving with the help of pick up and drop up probabilities.

Entropy:

The quality of cluster analyzed using two measures: Entropy, F-Measure. For each cluster, the class distribution of the data is calculated first.

$$E_j = -\sum_i p_{ij} \log(p_{ij)} \qquad (3.3.1)$$

Whenever the ant crosses a cluster, it will pick an item from the cluster and drop it into another cluster while moving.

$$\text{Pickup probability } Pp = \left(\frac{k1}{k1+f}\right)^2 \qquad (3.3.2)$$

$$\text{Drop up probability } Pd = \left(\frac{f}{k2+f}\right)^2 \qquad (3.3.3)$$

where f -> Entropy value, (Calculated before the item pick up & drop up)
k1 , k2 -> Threshold Constants.

- If Pd<Pp , item is dropped into another cluster and entropy value calculated again.

3.4. AIS with clustering & classification:

AIS is try to imitate real immune system. Most AIS use only the main ideas of real immune systems, namely clonal and negative selections which deal with the evolution of B-cells and T-cells, respectively [3]. The proposed method belongs to the method derived from immune system paradigm, called Clonal-G Selection.

- It resembles the original K-Means algorithm, but it get rid of its main drawback
    - It is able to estimate the proper number of cluster & avoids getting stuck in inappropriate areas.
    - Comparing to other immune algorithms for data clustering, its computational cost is decreased by producing a limited number of clones and proper suppression mechanism.
- Clonal Selection is used in data compression, data and web mining, clustering and optimization. Traditional clustering algorithm is to find the data distribution using cluster centers. When used for classification, these cluster centers are sometimes not the best ones, especially when the number of cluster is too small. The Clonal selection aims to evade this obstacle by means of a new suppression mechanism, which focuses learning on the boundaries among classes.
- **Concept:**
    - B-cells with different receptors' shapes try to bind to antigens (Training & Testing data).
    - The best fitted B-cells become stimulated and start to proliferate and produce clones, which are then mutated at very high rates.
    - After this process is repeated, it will emerge better B-Cells (Best Solution).

Pseudo code of AIS Algorithm:

```
1. A set of antigens Ag is presented to the antibodies population Ab ;
2. The affinity measure  f of the antibodies in relation to the antigens is calculated;
3. The n highest affinity antibodies to the antigens are selected to be cloned,generating the antibody subset Ab {n} ;
4. The antibodies selected will be cloned according to their affinity to the antigens (as higher the affinity more clones it will generate), producing a C clones population;
5. The C clones population is subjected to an affinity maturation process at an inversely proportional rate to the affinity of the clone (as higher the affinity, lower the mutation rate),  and a new population of clones C* is produced;
6. The C* clones population is evaluated and its affinity measure f* in relation to the antigens is calculated;
7. The n matured antibodies of the highest affinity are selected to compose the next population generation, since its affinity is greater than its original antibodies;
8. The d worst antibodies are removed from the population and replaced by new randomly generated antibodies.
This process repeats until a stop condition (number of generations) is reached.
```

- The population of B-cells depends on several mechanisms
    - Recognition
    - Stimulation
    - Proliferation
    - Hyper mutation
    - Suppression

Recognition:

Recognition of antigens (Training & Testing data) by B-cells depends on level of binding between them.

Stimulation:

The level of binding can be stimulated by given metric (Eg: Euclidean distance).

Hyper mutation:

It can be easily done by random changes in feature vector describing B-cells.

Suppression:

It is playing a vital role in the processing.

Clonal-G Selection:

- Simple
- Idea:
    - Clone generation (Choose number of attributes)
    - Suppression mechanism (Remove useless cluster's center).

4. RESULT AND DISCUSSION

4.1 GAC and MAC Empirical studies

GAC was executed with varied parameters values on the Breast database in order to choose the best setting among the existing possibilities, such as, mutation operator, crossover operator, mutation and crossover rates [1].

In these experiments the following data were analyzed: (a) Best solution; (b) Worst solution; (c) Average of the best solutions; (d) Standard deviation; (e) Average number of objective function evaluations.

Table 4.1 K-Means with Genetic

Database	**Algorithm**	**Fitness of the Best Solution**	**Fitness of the Worst Solution**	**Average Fitness of the Solutions**	**Standard Deviation**	**Average of the Fitness Evaluations**
Breast	ACO	334.00	334.00	335.00	0.90	23,060.00

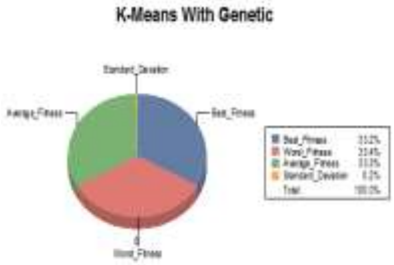


Figure-4.1: Graph result of GAC

4.2 ACO Empirical Studies

- Ants are used to cluster the data points using entropy [1].
- Entropy compute the probability (Pi,j) that the member of cluster belongs to the class.

Table 4.2 K-Means with ACO

Database	**Algorithm**	**Fitness of the Best Solution**	**Fitness of the Worst Solution**	**Average Fitness of the Solutions**	**Standard Deviation**	**Average of the Fitness Evaluations**
Breast	ACO	334.00	334.00	335.00	0.90	23,060.00

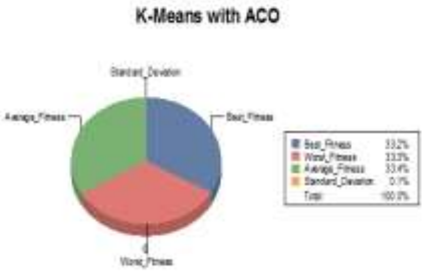


Figure-4.2: Graph result of ACO

4.3 AIS Empirical Studies

- The algorithms were executed with varied input parameter values on the breast database [1].
- In CLONLG algorithm, number of cluster, number of attributes to choose will be defined.

Database	**Algorithm**	**Fitness of the Best Solution**	**Fitness of the Worst Solution**	**Average Fitness of the Solutions**	**Standard Deviation**	**Average of the Fitness Evaluations**
Breast	AIS	383.00	1,194.00	764.00	18.5	23,189.00

Table: 4.3 K-Means with AIS

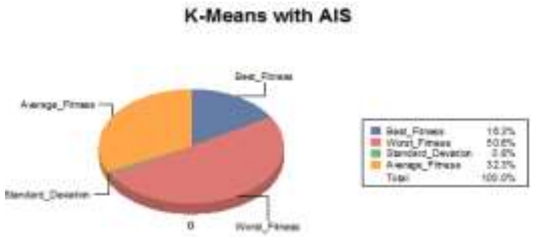


Figure-4.3: Graph result of AIS

4.4 ID3 Empirical Studies

- ID3 is a recursive process used to construct decision tree from data.
- Classification Trees are methodologies to classify data into discrete ones using the tree-structured algorithms. The main purpose of decision tree is to expose the structural information contained in the data.
- If the target variable (also called as response variable or class variable) is nominal/categorical variable is called "classification tree" and if continuous, the tree is called "regression tree" .

Building decision tree:

- Calculate the entropy for every attribute using the data set S.
- Split the set S into subsets using the attribute for which entropy is minimum. Make a decision tree node containing that attribute.
- Recurse on subsets using remaining attributes. Stop splitting when all examples at a node have the same labels.

- Entropy
  - used to measure the uncertainty associated with a random variable
  - Entropy(S) = $\sum_{i=1}^{n} -p_i \log(p_i)$

- Information gain
  - Information gain is based on the decrease in entropy after a dataset is split on an attribute.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Value(A)} \left|\frac{S_v}{S}\right| Entropy(S_v)$$

(4.4.1)

The performance of each algorithm is measured with the best, worst and normal fitness values count of each algorithm with the Euclidean distance measure for searching process. The Breast cancer dataset measure the standard deviation and mean average distance value of algorithm finally proposed AIS with ID3 shows the best classification result than the existing GA with ID3 and ACO with ID3 algorithm in breast cancer classification with fitness values.

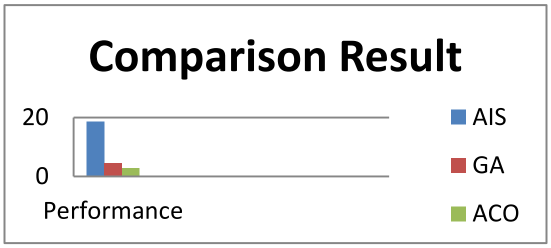
5. COMPARISON RESULT

In this section we measure the performance of the system and show the results of the accuracy in terms of how clustering performance is improved. The Breast cancer dataset measure the standard deviation and mean average distance value of algorithm finally proposed AIS with ID3 shows the best classification result than the existing GA with ID3 and ACO with ID3 algorithm in breast cancer classification with fitness values.

Table 5.1 Comparison result of ACO,GA,AIS

Database	Algorithm	Fitness of the Best Solution	Standard Deviation	Average of the Fitness Evaluations
Breast	GA	337.00	1.70	23,330.00
	ACO	334.00	0.90	23,060.00
	AIS	383.00	18.5	23,189.00

Figure 5.1 Comparison result of ACO,GA,AIS



6. CONCLUSION

In this work analyzed the performance of the GA, ACO and AIS metaheuristic algorithms with ID3 for solving data Clustering & Classification problem in an experiment with breast cancer dataset data of UCI repository. Instead of K-Mean clustering, this research work can combine the simplicity of the K-Means algorithm with the robustness of AGA-Miner.

In the proposed system, an efficient ID3 classification techniques are used to measure the performance of the system with GA, ACO and AIS system, which could increase the accuracy and reduces the cost of time. Proposed AIS system produces the best classification result than the ACO and GA based decision tree ID3 classifiers.

6.1 FUTURE ENHANCEMENT

As per our observation there are some future suggestions which are listed below:

- This experimental result may be used to detect other cancer types such as lung cancer, mouth caner and etc.
- Apply the other classification algorithm such as C4.5 and C5.0, Ripper classification algorithm and compare the results with ID3 methods.

Instead of applying the optimization methods different correlation based similarity measures with optimization are applied to cluster the dataset.

7. REFERENCES

[1] Ana Cristina B.Kochem Vendramin, Diogo Augusto Barros Pereira, “Application of Bio-inspired Metaheuristics in the data clustering problem”

[2] Benny Pinkas, Yehuda Lindell, “Privacy Preserving Data Mining”.

[3] Berkhin, P. 2002. “Survey clustering data mining techniques”, Technical report, Accrue software, San Jose, California.

[4] M. Bramer. “Principles of Data Mining”.Springer, 2007.

[5] D. Dasgupta, Artificial Immune Systems and Their Applications, Springer, Berlin, 1999 .

[6] Ding, C., and He, X. 2002. “Cluster merging and splitting in hierarchical clustering algorithms”, IEEE international conference, pp. 139-146.

[7] Dorigo, M., Maniezzo.V,& Colorni .A.,” Ant System: Optimization by a colony of cooperating agents,” IEEE Transactions on Systems, Man, and Cybernetics – Part B, 26, 29–41,1996.

[8] G. Garai and B. B. Chaudhuri. “A Novel Genetic Algorithm for Automatic Clustering”. Pattern Recognition Letters, vol. 25, n.2, pp.173-187, 2004.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 2, 100 - 104, 2014

Insuring Security for Outsourced Data Stored in Cloud Environment

Durga Priya.G
Department of Information Technology
Sri Sairam Engineering College
Chennai-45, India

Soma Prathibha
Department of Information Technology,
Sri Sairam Engineering College
Chennai-45, India

Abstract -- The *cloud* storage offers users with infrastructure flexibility, faster deployment of applications and data, cost control, adaptation of cloud resources to real needs, improved productivity, etc. Inspite of these advantageous factors, there are several deterrents to the widespread adoption of cloud computing remain. Among them, security towards the correctness of the outsourced data and issues of privacy lead a major role. In order to avoid security risk for the outsourced data, we propose the dynamic audit services that enables integrity verification of untrusted and outsourced storages. An interactive proof system (IPS) with the zero knowledge property is introduced to provide public auditability without downloading raw data and protect privacy of the data. In the proposed system data owner stores the large number of data in cloud after e encrypting the data with private key and also send public key to third party auditor (TPA) for auditing purpose. TPA in clouds and it's maintained by CSP. An Authorized Application (AA), which holds a data owners secret key (sk) and manipulate the outsourced data and update the associated IHT stored in TPA. Finally Cloud users access the services through the AA. Our system also provides secure auditing while the data owner outsourcing the data in the cloud. And after performing auditing operations, security solutions are enhanced for the purpose of detecting malicious users with the help of Certificate Authority.

Keywords : Data Security, Certificate Authority, Audit service, Cloud storage, Dynamic operations, Data verification.

1. INTRODUCTION

CLOUD Computing is generally a virtual servers available over the Internet.According to NIST[14], CLOUD computing can be defined as "It is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

This cloud model is composed of five essential characteristics, three service models, and four deployment models." Services[14] of cloud computing are 1.SaaS(Software as a Service), 2.PaaS(Platform as a Service), and 3.IaaS (Infrastructure as a Service). *SaaS*: run on distant computers "in the cloud" that are owned and operated by others and that connect to users' computers via the Internet and, usually, a web browser. *PaaS*: provides a cloud-based environment with everything required to support the complete life cycle of building and delivering web-based (cloud) applications—without the cost and complexity of buying and managing the underlying hardware, software, provisioning and hosting.

IaaS: provides companies with computing resources including servers, networking, storage, and data centre space on a pay-per-use basis.

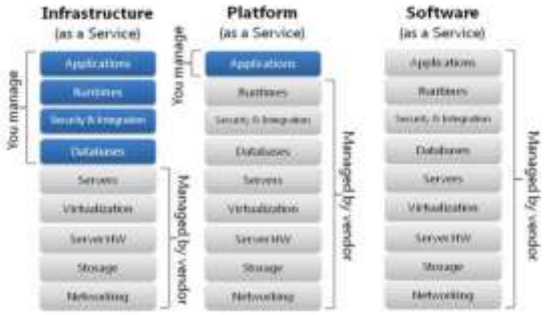


Figure 1: Cloud Services[13]

Cloud Computing has three Deployment models. They are Public cloud, Private cloud, and Hybrid cloud. *Public clouds* are owned by companies and users don't need to purchase hardware, software which are owned and managed by providers. *Private clouds* are owned by single company and take advantage of many of cloud's efficiencies, while providing more control of resources and steering clear of multi-tenancy. *Hybrid cloud:* uses a private cloud foundation combined with the strategic use of public cloud services.

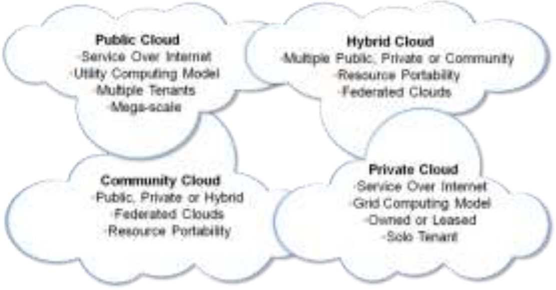


Figure 2: Deployment models[13]

Among these deployment strategies, the public clouds face a huge drawback, which is called to be as security issue. The issues are threats to data, loss of data integrity, confidentiality, and reliability and so on. These are the hindrances that stop the growth of cloud computing technology. This occurs in public cloud because, entities like Cloud Service Provider (CSP), Third Party Auditor (TPA) are involved which may act disloyally to the data's in the cloud. For example, to increase the profit margin by reducing cost, it is possible for CSP to discard rarely accessed data without being detected in a timely fashion. Similarly, CSP may even attempt to hide data loss incidents so as to maintain a reputation. Therefore, although outsourcing data into the cloud is economically for the cost and complexity of long-term large-scale data storage, it's lacking of offering strong assurance of data integrity and availability may impede its wide adoption by both enterprise and individual cloud user. In order to achieve the assurances of cloud data integrity and availability and enforce the quality of cloud storage service, efficient methods that enable on-demand data correctness verification on behalf of cloud users have to be designed. However, the fact that users no longer have physical possession of data in the cloud prohibits the direct adoption of traditional cryptographic primitives for the purpose of data integrity protection. Hence, the verification of cloud storage correctness must be conducted without explicit knowledge of the whole data files. The cloud computing is deployed by data centres running in a simultaneous, cooperated and distributed manner. Hence ensuring security for outsourced data in cloud is the most important task of all.

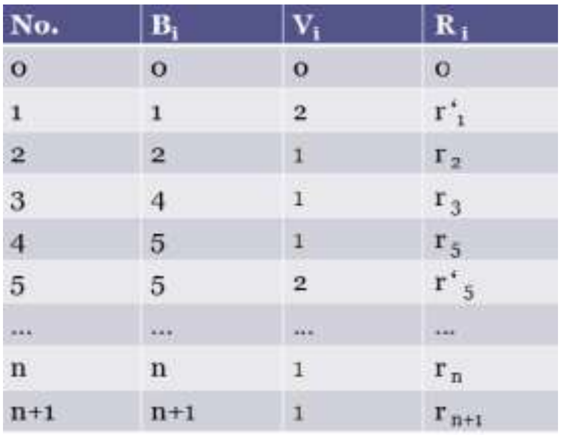
2.RELATED WORK

Many mechanisms have been proposed to ensure the security of cloud users and for their data. Yet once the malicious users acquire the security credentials they can pose as genuine users and hack the data**.** In this section, will discuss about the work carried out in the area of cloud security. In [1] Dynamic audit sources provide the user with performance of the audit services but it doesn't make an effort to verify if the user is genuine or not. Though privacy preserving [2] works more efficiently but it's only for encrypted files. The effectiveness of this lies in the hands of auditors, whose statefulness must not affect it, also the limited number of auditors matter. A random spot verification mechanism is developed in [3], which correctly identifies where there has been modification and it is efficiently resilient to changes and malicious attacks. But the inefficiency largely attractive is attributed to the randomness. If an identity based mechanism is used [4], can avoid key revocation and key escrow problems but other types of problems are not concentrated here. While building PDP technique on [6] symmetric key, it will considerably reduce the cost and bulk encryption but it is not very safe when it comes to public users. And can considerably try to provide security by auditing the data [7] which is inserted, it overloads the client side as all the auditing is done there. Hence this is useful for smaller data insertions. This can also be implemented by creating probabilistic proofs of data possessed[8] , this way the user can be sure of the data he has uploaded and the data that has been retrieved. This approach's efficiency is reduced by the large volume of data loaded and verified.

3.PROPOSED WORK

Audit service is constructed based on the techniques, fragment structure[1], random sampling[1], and index-hash table[1], supporting provable updates to outsourced data and timely anomaly detection. Also propose a method based on probabilistic query and periodic verification for improving the performance of audit services. Security solutions also introduced to avoid the malicious users while outsourcing in the cloud. Audit system can support dynamic data operations[1] and timely anomaly detection[1]. Security is provided for dynamic data operations and detects the malicious cloud service provider, when accessing the data in the cloud. We also Detect the malicious identity while the data owner outsourcing in the cloud. First the data centres are configured and then while outsourcing the data onto cloud, authentication for data owner is performed. After performing this verification, a file that has to be uploaded is chosen. From the selected file, we generate Public Verifiable Parameters (PVP), Index Hash Table (IHT), and Tags. PVP and IHT are sent to TPA and Tags are sent to CSP for security purpose. Once the data owner uploads the file in the cloud, the TPA is checking the integrity of the uploaded file at any time.

Table 1: Index Hash Table[1]



No.	B_i	V_i	R_i
0	0	0	0
1	1	2	r'_1
2	2	1	r_2
3	4	1	r_3
4	5	1	r_5
5	5	2	r'_5
...	...	...	...
n	n	1	r_n
n+1	n+1	1	r_{n+1}

Thus the 3-move interactive proof protocol is used among the TPA and cloud service provider for the auditing purpose. 3- Move interactive protocols are commitment, challenge and response.

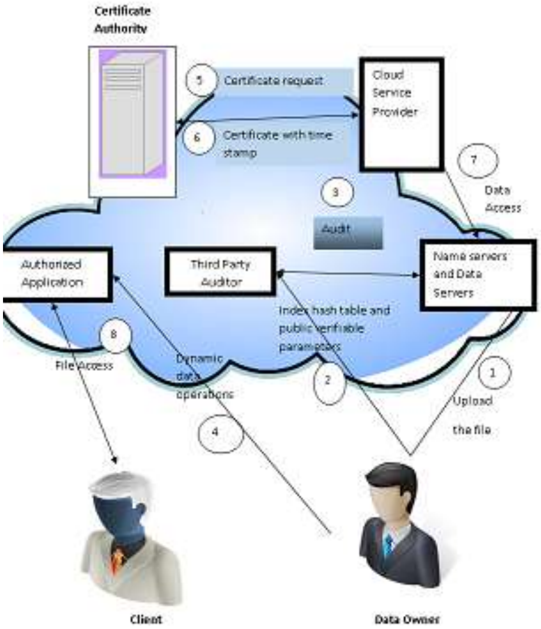


Figure 3: System Architecture

At first the TPA queries the CSP for the verification process and initializes the interactive proof protocol. The cloud service provider selects some set of random keys and random blocks and sent it the TPA using the commitment protocol. Next the TPA chooses some set of secret keys and blocks and sends to the CSP by using the challenge protocol. After which cloud service provider calculates the response and send to the TPA. The verifier TPA checks whether the response is correct. By doing so the auditing is performed among the CSP and TPA.

3.1 KeyGen(1^k)[1]:

1.Bilinear map group system=(p,G,G_T ,e)
2.Collision resistant hash function= H_k
3.chooses a random α,β Є $_R Z_p$ and computes $H_1 = h^\alpha$ and $H_2 = h^\beta$ Є G.

3.2 TagGen(sk,F)[1]:

1.Splits the file F into n×s sectors
2.chooses s random $\tau_{1,} \dots \tau_s$ Є Z_p (secret of the file)
3.computes $u_i = g^{\tau i}$ Є G and $\xi^{(1)} = H_\xi$ ("Fn")
4.where $\xi = \Sigma^s{}_{i=1}\ \tau_i$ and Fn is the filename.
5.Finally sets $u=(\xi^{(1),}\ u_1 ,\dots, u_s)$ and outputs $\psi=(u, \chi)$ where χ is the index hash table.

4.IMPLEMENTATION

This system is implemented in CLOUD ANALYST. CloudAnalyst[14] is a framework which enables seamless modelling, simulation and experimenting on designing Cloud computing infrastructures.

CloudAnalyst is a self-contained platform which can be used to model data centres, service brokers, scheduling and allocation policies of a large scaled Cloud platform.

It provides a virtualization engine with extensive features for modelling the creation and life cycle management of virtual engines in a data centre. The CloudAnalyst is built directly on top of CloudSim framework leveraging the features of the original framework and extending some of the capabilities of CloudSim.

The modules are developed using Java in JCreator which is a Java IDE. This interface is similar to that of Microsoft Visual Studio.

a. It depicts the configuration of data centres.

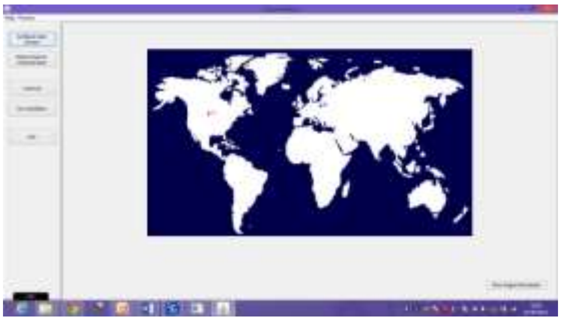
b. Data Owners identity is authenticated with the help of secret key and public key.

c. Index Hash Table and Public Verifiable Parameters are generated and sent to Third Party Auditor

d. Commitment is performed between CSP and TPA. This action is initiated by CSP

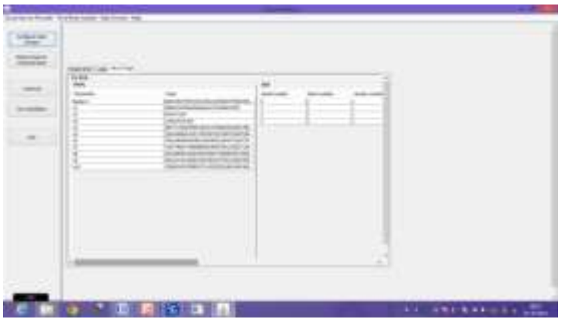
e. Response is sent to CSP from TPA.

f. After performing Check operation, Auditing was completed.



a. Data Centres in Cloud Analyst

b. Authentication of Data Owner



c. Generating PVP and IHT

d. Commitment- Auditing Services

e. Response- Auditing Services

f. Check - Auditing Services

5. CONCLUSION

Outsourcing has become critical to business operations and vital for businesses to sustain their competitive advantages. Maintaining security in IT outsourcing is important for maintaining the growth of IT outsource services. Thus proposed approach provides the security outsourcing services by enabling periodic audit and dynamic operations. Also the verification is provided for the cloud service provider to access the data in the cloud. Hence the malicious cloud service providers are removed from the system.

6.REFERENCES

[1] Yan Zhu,Gail-Joon Ahn, Hongxin Hu, Stephen S. Yau, Ho G.An and Chang-Jun Hu" Dynamic Audit Services for Outsourced Storages in Clouds."

[2] M. Mowbray, "The Fog over the Grimpen Mire: Cloud Computing and the Law," Technical Report HPL-2009-99, HP Lab., 2009.

[3] A.A. Yavuz and P. Ning, "BAF: An Efficient Publicly Verifiable Secure Audit Logging Scheme for Distributed Systems," Proc. Ann.Computer Security Applications Conf. (ACSAC), pp. 219-228, 2009.

[4] G. Ateniese, R.C. Burns, R. Curtmola, J. Herring, L. Kissner, Z.N.J. Peterson, and D.X. Song, "Provable Data Possession at Untrusted Stores," Proc. 14th ACM Conf. Computer and Comm. Security, pp. 598-609, 2007.

[5] G. Ateniese, R.D. Pietro, L.V. Mancini, and G. Tsudik, "Scalable and Efficient Provable Data Possession," Proc. Fourth Int'l Conf. Security and Privacy in Comm. Netowrks (SecureComm), pp. 1-10, 2008.

[6] C.C. Erway, A. Ku¨ pc¸u¨ , C. Papamanthou, and R. Tamassia, "Dynamic Provable Data Possession," Proc. 16th ACM Conf. Computer and Comm. Security, pp. 213-222, 2009.

[7] H. Shacham and B. Waters, "Compact Proofs of Retrievability," Proc. 14th Int'l Conf. Theory and Application of Cryptology and Information Security: Advances in Cryptology Advances in Cryptology (ASIACRYPT '08), J. Pieprzyk, ed., pp. 90-107, 2008.

[8] H.-C. Hsiao, Y.-H. Lin, A. Studer, C. Studer, K.-H. Wang, H. Kikuchi, A. Perrig, H.-M. Sun, and B.-Y. Yang "A Study of User- Friendly Hash Comparison Schemes" Proc. Ann. Computer Security Applications Conf. (ACSAC), pp. 105-114,2009.

[9] A.R. Yumerefendi and J.S. Chase, "Strong Accountability for Network Storage," Proc. Sixth USENIX Conf. File and Storage Technologies (FAST), pp. 77-92, 2007.

[10] Y. Zhu, H. Wang, Z. Hu, G.-J. Ahn, H. Hu, and S.S. Yau, "Efficient Provable Data Possession for Hybrid Clouds," Proc. 17th ACM Conf. Computer and Comm. Security, pp. 756-758,2010.

[11] M. Xie, H. Wang, J. Yin and X. Meng, "Integrity Auditing of Outsourced Data" Proc. 33rd Int'l Conf. Very Large Databases (VLDB), pp.782-793,2007.

[12] C. Wang, Q. Wang, K. Ren and W. Lou, "Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing" Proc. IEEE INFOCOM, pp. 1-9, 2010.

[13] www.googleimages.com

[14] www.wikipedia.com

A Review of Detection and Reduction of Noise in Degraded Images by Efficient Noise Detection Algorithm

Gayatri P. Bhelke
Sipna college of engineering and technology, Amravati
Akola, India

V. S. Gulhane
Sipna college of engineering and technology, Amravati
Amravati , India

N. D. Shelokar
Sipna college of engineering and technology, Amravati
Amravati , India

Abstract: Noise is unwanted information present in image that can harm the quality of image. A novel noise detection and reduction technique is proposed in this paper. Technique is used to detect and reduce the noise in the degraded images. The main approach of this proposed technique is detecting and reducing different types of noise present in degraded images. Paper introduces a two step process where first step is used to detect the type of noise which is present in degraded image and second step handles the errors and compensate for various noise sources common in multimedia application, such as Salt and Pepper, Gaussian and speckle noise.

Keywords: Image Restoration, Degraded Images, Denoising , Noise Detection , Salt and pepper Noise, Gaussian .

1. INTRODUCTION

Noise is unwanted information which is present in the images, which affect the quality of images. Noise can be unavoidable in communication networks, and its presence can have terrible effects upon the data being sent [1]. Image Processing is a technique that improves the quality of raw Images capture in normal day-to-day life for many applications. Images captured by digital cameras could be affected by noise due to random variations of pixel elements in the camera sensors. There are various types of image noise present in the image such as Gaussian noise, salt and pepper noise, random valued impulse noise, speckle noise, Uniform noise [4].

- **Salt And Pepper Noise**

Salt and Pepper noise is also known as Impulse Noise. This noise can be caused by sharp & sudden disturbances in the image signal. It represents itself is randomly occurring white or black (or both) dots over image.

- **Gaussian Noise**

Gaussian Noise is caused by random fluctuations in the signal.

Its modeled by random values added to an image

- **Speckle Noise**

Speckle noise can be modeled by random values multiplied by pixel values of an image.

- **Uniform Noise**

Uniform noise is also known as quantization noise. It is caused by quantizing the pixels of a sensed image to a number of discrete levels. It has an approximately uniform distribution

In this work, we will present a new, faster and more efficient noise detection and reduction method for degraded images. The algorithm which is used to detect the presence of noise and to remove it, should be theoretically, and computationally as simple as possible. A well-defined process of detecting and reducing certain types of noise in transmitted images would have to be a somewhat crude for speed. The main aim of noise reduction is to smother the noise, and also probably to safeguard the harpness of edge and feature information [23].

Here, we have discussed a new efficient noise detection algorithm that will be able to allow the receiver to know what type of filtering method should be applied for the type of noise detected in given image.

2. RELATED WORK

Many of the current papers dealing with noise in communication networks which propose a two-stage method of impulse noise reduction where in the first stage noise is detected and in the second it compensated for a filtering technique.

As per Ming Yan’s paper [10] when the noise level is not high, adaptive center-weighted median filter (ACWMF) is appropriate method for removing random-valued impulse noise. Paper presents a general algorithm for blind image inpainting and removing impulse noise by iteratively restoring the image and identifying the damaged pixels.

In H. Hosseini, F. Marvasti’s [5] paper GFN(General Fixed-Valued Impulse Noise) model is used. GFN model required an Impulse Value Detector (IVD) to determine the noise values. In this received image is denoted as I and Image entrophy is defined as,

$$\text{entropy}(\mathbf{I}) = \sum_{i=0}^{255} \mathbf{pi} \log \mathbf{pi}$$

where, pi is the probability of the grey-value i and can be interpreted as the normalized histogram of the image. While the Gaussian noise does not affect the image entropy, the impulse noise significantly decreases it.

The impulse value detector, iteratively, detects and removes the impulse grey-values. If the corresponding pixels have the lowest correlation with their neighbors then the grey value is detected as an impulse. After each iteration, when the impulse value is removed, the image entropy increases. The process continues until the entropy becomes larger than the entropy threshold, thus it ensures that there are no more impulse values in the image and image restoration is done by using AIM filter. In this filter, the noisy pixels which are farther than their nearest uncorrupted pixel, will be modified in more iterations

Qin Zhiyuan et.al ‘s[14] describe A Robust Adaptive Image Smoothing Algorithm in which analysis of some smoothing algorithms are given which include Edge Preserved filtering ,Adaptive medium filter, Robust smoothing filter and Gradient weighting filter. Robust adaptive algorithm combines multi-window templates, gradient weighting, constant gray output on non-pulse pixel and the improved adaptive smoothing algorithm.

In Addition to Smoothing algorithm, paper introduces the methods of enlarging windows and selecting sub template windows to remove salt and pepper noise with large space intensity. Because it uses a new algorithm by combining nonlinear filtering and linear filtering according to their respective adaptation to different noises.

In Deborah D. Duran-Herrmann et.al [1] paper two stage process is given in which first stage detect the type of noise and in second stage which type of filter is suitable for detected noise is given to eliminate the noise.

“Noise removal Algoritham for Image Corrupted by Additive Gaussian Noise”[20] describe two fundamental mathematical morphological operations , that are dilation and erosion. Dilation adds pixels to the boundaries of objects in an image, where as erosion removes pixels on object boundaries. Mathematical morphological operations are also useful in smoothing and sharpening. Paper presents noise removal algorithm for gray scale images corrupted by Gaussian noise.

3. PROPOSED WORK

Our given algorithm uses a two-stage process of determining three things which are given below,

- Presence of noise
- Type of noise such as impulse noise or Gaussian
- The effective filtering method for removing noise

Similar to [1], architecture includes Adaptive Noise Detector and Adaptive pixel Restorer

Following fig shows system architecture,

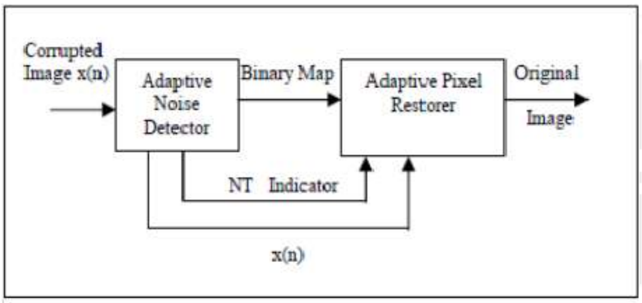


Figure. 1 . System Architecture

Following subsection describe Adaptive Noise Detector and Adaptive pixel Restorer.

3.1 Adaptive Noise Detector

The Adaptive Noise Detector is used to detect the type of noise such as Gaussian noise, salt and pepper and so on, if exists in the current image. Similar to [8] it follows the following steps,

Step 1: Obtain the image histogram H of the degraded image

Step2: Compute the vector *D* which is the difference between adjacent locations in the histogram array *H*.

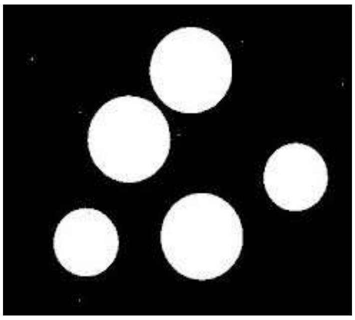
$Di = Hi+1 - Hi$ *for all i=0,1…255*

Step 3: Various boundary thresholds are set, according to the maxima values found in D. Depending on location of maxima values nature of the noise is detected, whether it be impulse, uniform, or Gaussian. Once the noise has been find out, the NT Indicator is set as an input to the second stage of the system. The corrupted pixels are then mapped to a binary matrix with the same dimensions as the image

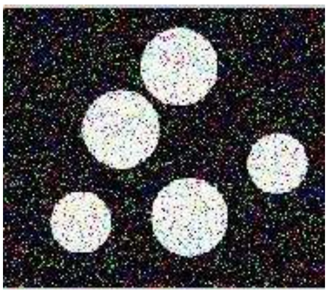
3.2 Adaptive pixel Restorer

In second stage NT indicator is used as input which increases the processing speed. If the NT Indicator has not been set, the “corrupted image”, x(n), is allowed to pass and if the NT Indicator is set, then this sets one of the various noise flags. Again, similar to the second stage of [1], the Adaptive Pixel Restorer searches through the Binary Map for pixels whose value is “1”. If the neighborhood vector of values is not set as null and instance is found , it searches noise flags that are set to determine what type of adaptive filtering is best for that instance which is courrpted. If the neighborhood vector of values in the vicinity of that particular pixel is null, the algorithm goes on to the next pixel value. This process is done repeatedly until the Binary Map is cleared of 1’s.[1]

Following fig(a) shows the original image fig(b) shows image corrupted by salt and pepper noise with histogram , fig(c) shows image corrupted by the Gaussian noise with histogram



Figure(a) original Image



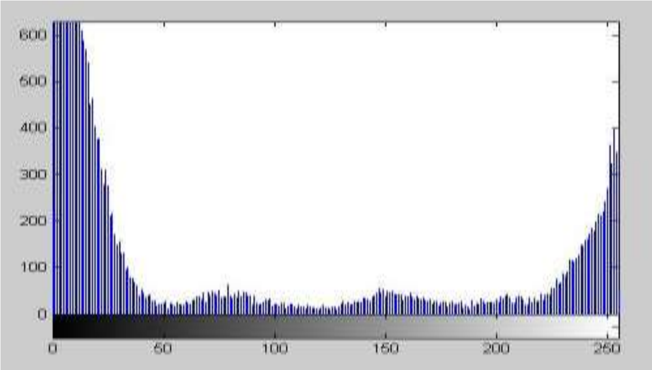


Figure (b) Image corrupted by Salt and pepper noise histogram of salt and pepper noise.

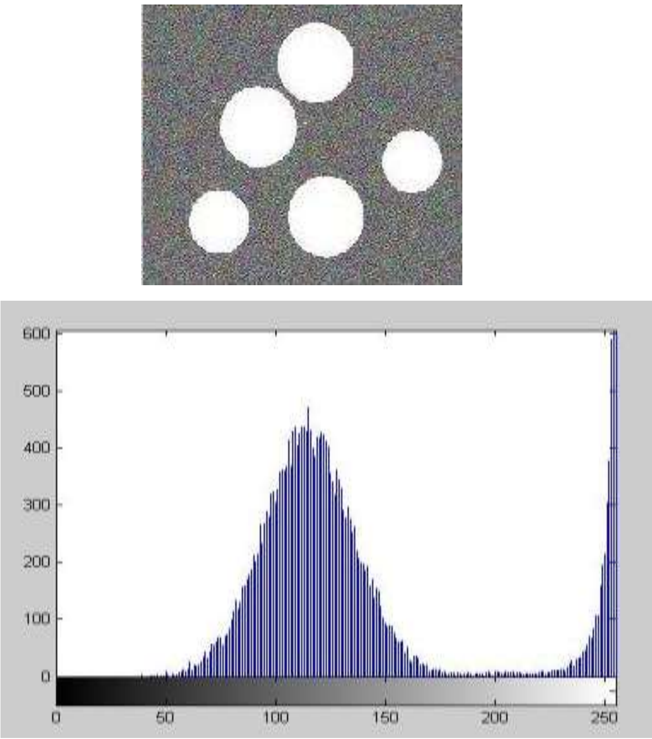


Figure (c). Image corrupted by Gaussian Noise with histogram

Objectives of this proposed work are summarized as follow

- Analysis of algorithm which used for detection of noise in corrupted images.
- Implementation of Adaptive/Novel Algorithm for detection and reduction of Noise in corrupted images.
- Implementation will be carried out in two step process,
  - ➢ Presence of noise
  - ➢ Type of noise
  - ➢ Effective filtering method for removing noise.

With the best level efforts, above one or more task may be tried to be implemented

4. REFERENCES

[1] [Deborah D. Duran-Herrmann et al 2012], Qilin Qi, and Yaoqing (Lamar) Yang "An Adaptive Algorithm for Corrupted Images" *International Conference on Systems and Informatics* (ICSAI 2012)

[2] [D. Ze-Feng et al 2007], Y. Zhou-Ping, and X. You-Lun "High Probability Impulse Noise- Removing Algorithm Based on Mathematical Morphology," *IEEE Signal Processing Letters*, Vol. 14, No. 1, January 2007, pp. 31-34.

[3] [Gouchol Pok et al 2003], Jyh-Chain Liu and A.S. Nair, "Selective removal of impulse noise based on homogeneity level information", *IEEE Trans on image processing*, vol.12, no.1, pp.85-92, January 2003.

[4] [Gurmeet kaur et al 2012] Rupindar Kaur "Image De-noising using Wavelet Transform and various filters*"*, *International journal of researcher in computer science* Eissn 2249-8265 volume 2(2012)pp.15-21W.

[5] [H. Hosseini et al 2011], Senior Member, IEEE "Fast Impulse Noise Removal from highly corrupted image"

[6] [How-Lung Eng et al 2001], Kai-Kuang Ma," Noise adaptive soft switching median filter", *IEEE Trans on Image process.*, vol.10, no.2, pp.242-251, Feb.2001.

[7] [Ho-Ming Lin et al 2007],Alan N Willson, "Median filters with adaptive Length", *IEEE Trans on circuits and systems*, vol.35,no.6, pp.675-690, June 1988.

[8] .[Indu S and Chaveli Ramesh 2007], "A Noise Fading Technique for Images Highly Corrupted with Impulse Noise" *International Conference on Computing: Theory and Applications (ICCTA'07).*

[9] [Luo, 2006], "Efficient Removal of Impulse Noise from Digital Images,"*IEEE Transactions on Consumer Electronics*, Vol. 52, No. 2, May 2006,pp. 523-527.

[10] [Ming yan 2011], "Restoration of Images Corrupted by Impulse Noise using Blind Inpainting and l0 Norm".

[11] [Pei-Eng Ng et al 2006], Kai-Kuang Ma, "A switching median filter with BDND for extremely corrupted images", *IEEE Trans on Image process*., vol. 15, no.6, pp.1506-1516, June 2006.

[12] [Peter D. Wendt et al 1986], Edward J Coyle and Neal. C. Gallagher, "Some convergence properties of median filters", *IEEE Transon circuits and systems*, vol. cas-33, no.3, pp.276-286, March 1986

[13] [Ping et al 2007], W.; Junli, L.; Lu, D.; and Chen, G. "A Fast and Reliable Switching Median Filter for Highly Corrupted Images by Impulse Noise," *IEEE Transactions*, 2007, pp. 3427-3440.

[14] [Qin Zhiyuan a et al 2010] Zhang Weiqiang a, Zhang Zhanmu a, Wu Bing b, RuiJie a,ZhuBaoshan "A Robust Adaptive image smoothing algorithm"

[15] [R. Gonzales et al 2002] R. Woods "Digital Image Processing" *Second Edition Prentice-Hall*, Inc. 2002.

[16] [Raymond H Chan et al 2005], Chung-Wa Ho and Mila Nikolova, "*Salt* nd *pepper* noise removal by median type noise detectors and detail preserving regularization",

IEEE Trans on Image Processing, vol.14, no.10, pp. 1479-1485, October 2005.

[17] [Reinhard Berstein 1987], "Adaptive nonlinear filters for simultaneous removal of different kinds of noise in images",*IEEE Trans on circuits and systems*, vol.Cas-34, no.11,pp.1275-1291, November 1987

[18] [S. Schulte, et al 2006] M. Nachtegael, V. De Witte, D. Van der Weken, E.E.Kerre, "A Fuzzy Impulse Noise Detection and Reduction Method,"*IEEETransactions on Image Processing*, Vol 15, Issue 5, May 2006, pp.1153 - 1162.

[19] [S. Schulte, M. Nachtegael et al 2006] V. De Witte, D. Van der Weken, E.E. Kerre, "Fuzzy Two-Step Filter for Impulse Noise Reduction From Color Images,"*IEEE Transactions on Image Processing*" Vol 15, No 11, November 2006, pp.3568-3579.

[20] [Shyam Lall et al 2009] Mahesh Chandra, Gopal Krishna Upadhya "Noise Removal Algoritham for Image Corrupted by Additive Gaussian Noise",*International Journal of Recent trends in Engineering*,vol 2,No.1,November 2009,pp199-206

[21] [T. Chen et al 2006] S. Huang, C. Chen, and Z. Lin, "Adaptive Working Window for Impulse Noise Reduction," *International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2006.*

[22] [Tao Chen et al 1999], Kai-Kuang Ma , Li- Hui- Chen, "Tristate median filter for image denoising", *IEEE Trans on image processing*, vol.8, no.12, pp.1834-1838, December 1999.

[23] [V. Saradhadevi et. al 2011], Dr.V.Sundaram, "An Enhanced Two-Stage Impulse Noise Removal Technique based on Fast ANFIS and Fuzzy Decision", *IJCSI International Journal of Computer Science Issues*, Vol. 8, Issue 5, No 1, September 2011.

[24] [W. Luo et al 2006] D. Dang, "An Efficient Method for the Removal of Impulse Noise," *ICIP*, 2006, pp. 2601-2604

[25] [W. Luo,2006] "An Efficient Detail-Preserving Approach for Removing Impulse Noise in Images," *In IEEE Siganl Processing Letters*, Vol 13, No 7, July 2006, pp. 413-416.

[26] [W. Luo,2006] "Efficient Removal of Impulse Noise from Digital Images,"*IEEE Transactions on Consumer Electronics*, Vol. 52, No. 2, May 2006,pp. 523-527

[27] [Xiaoyin Xu et al 2004], Eric L Miller, Dong bin Chen and Mansoor Sarhadi, "Adaptive two-pass rank order filter to remove impulse noise in highly corrupted images", *IEEE Trans on image processing*, vol.13, no.2, pp. 238-247, February 2004

[28] [Xu, H et al 2006], Xia, X.; Guo, L.; Chen, W.; and Huang, G. "Classification-based Weighted Filter for Image Corrupted by Impulse Noise," *ICSP Proceedings, 2006*.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 2, 109 - 113, 2014

A New Method for Reducing Energy Consumption in Wireless Sensor Networks using Fuzzy Clustering and Fault Tolerance

Vahab ashrafian
Department of Computer Science and Research Branch
Islamic Azad University
Ardabil, Iran

Ali Harounabadi
Member of Science Board of Computer Group in Azad Islamic university of Tehran Center, Iran

Mehdi Sadeghzadeh
Member of Science Board of Computer group in Azad Islamic university of Mahshahr, Iran

Abstract: Nowadays, wireless sensor networks, clustering protocol based on the neighboring nodes into separate clusters and fault tolerance for each cluster exists for sensors to send information to the base station, to gain the best performance in terms of increased longevity and maintain tolerance than with other routing methods. However, most clustering protocols proposed so far, only geographical proximity (neighboring) cluster formation is considered as a parameter. In this study, a new clustering protocol and fault tolerance based on the fuzzy algorithms are able to clustering nodes in sensor networks based on fuzzy logic and fault tolerance. This protocol uses clustering sensor nodes and fault tolerance exist in the network to reduce energy consumption, so that faulty sensors from neighboring nodes are used to cover the errors, work based on the most criteria overlay neighbor sensors with defective sensors, distance neighbor sensors from fault sensor and distance neighbor sensors from central station is done. Superior performance of the protocol can be seen in terms of increasing the network lifetime and maintain the best network tolerance in comparison with previous protocols such as LEACH in the simulation results.

Keywords: Wireless sensor networks; fault tolerance; fuzzy algorithms; clustering; faulty sensor.

1. INTRODUCTION

Nowadays, remote control and monitoring systems are one of the challenging issues in the field of electronics and computer science. This investigation whenever looking for a solution to take into accounts the specific conditions and expectations for the answer. In terms of quality and the same thing, whatever the cost /effectiveness ratio is lower, the same way as its popularity is higher. Aware of their surroundings or changes the state of each set of equipment that can be used as sensors are known. Each sensor can take into account the changing environment in terms of certain parameters such as temperature, humidity, pressure ... Senses and offer. Based on information from a set of sensors embedded in the environment, can controls environment and its changes.
Recent advances in electronics and wireless sensors make it possible multiple-purpose sensors provided with low energy consumption and cost. These sensors are able to communicate with each other over short distances. A node is a very small sensor with sensing equipment, data processing and wireless communication. In fact, a sensor network is a collection of numerous sensor nodes are distributed in the environment and each autonomously and in collaboration with other nodes to follow a particular purpose. Nodes are close together, and each node can communicate with another node and other nodes located in their data available to the environment, finally the status of desired environment to be reported to central point.

Techniques and methods used in sensor networks are highly dependent on the nature of the application, network topology structure, and environmental conditions, limits and efficiency and cost parameters. So today, throughout the reputable universities and computer research centers, electronics and telecommunications, wireless sensor networks are considered as a very attractive and popular research field. Many suggestions and research on various topics sensor network has been presented and volume of research in this field is increasing. The main goal of all these efforts and provide solutions, having a system with a simple control methods, which is easy and low cost to meet the desired needs (bandwidth, energy, and environmental interventions . . .) that could stood up against constraints and provide general conditions in accordance with the wishes and aspirations (transfer bulk data content, continuity, long life, low cost, etc.).

In order to neither do tasks in a wireless sensor network must consider the time and consumption energy to loss duly works nor drastic decline in lifetime of the network. In other words, the constraints facing these networks, and how much and manner of consume energy is of particular importance because complete loss of battery in the sensors means they are destroy and due to the environment and use networks changing of thousands of sensor node batteries is virtually impossible.
The reduction of energy consumption in wireless sensor networks has a direct relationship with increased longevity. Ideal mood in the sensor networks is that the energy of all nodes to finish concurrent and it is a mood in which highest lifetime is possible for the network. Thus, to increase in network lifetime trying to distribute the load on the network is a uniform distribution to mineralize the time between death of the first node and the last node.

To achieve this goal, several communication protocols have been proposed so far, in which the protocols are based on clustering, significantly lowers energy consumption.

The Protocols of the entire network is partitioned into several cluster-ware and each cluster has a node is selected as cluster heads. In this case, instead of each node sends its data directly to the base station, send it to the cluster heads, finally, the cluster heads then collecting and combining data from all of the nodes in the cluster, and data are sent to the central station.

In this protocol, choose a sensor as a replace faulty cluster and appropriate clustering significantly increase the network lifetime, scalability and efficiency.

2. Previous Work

Ahmadinia and his colleagues are provided a clustering method of nodes in sensor networks based on ICLA. In the clustering method, various parameters such as balancing the cluster size, clusters energy and ... taken into consideration and compared with other clustering methods, create clusters with more balanced and increases network lifetime [1].

Akbarzadeh and his colleagues declared that was originally selected cluster heads by fuzzy logic and considering the energy parameters of the sensor, numbers of neighbor sensor and spacing parameters for proper distribution of cluster heads done in the network then optimize point to move the base station using a genetic algorithm are determined to energy efficiency consumption of cluster heads [2].

Attention to the problem of clustering based routing in wireless sensor networks aimed to reduce energy consumption and maintain network coverage has been paid. For this purpose, neural network self-organizing map (SOM) is used to provide energy-based clustering protocols. The new protocol based on the energy of the self-organizing clustering protocol (EBCS) is called clustering according to three criteria: level of energy and spatial coordinates of each node are performed and its superiority in terms of longevity and maintains network covering (simulation) are proved. (Supervisor: Dr. Reza Askari Moghadam), (Supervisor: Dr. A. Taraghi Haghighat) [3].

Roshan Zadeh et al are provided an optimize algorithm for energy consumption and send the matched packets. In the proposed algorithm is called (PT-Multipath) routing decisions make is based on the residual energy of nodules. Also, select nodes for racing with information node one to one and double jumping node were sender. Simulation results show that the proposed algorithm efficiently distributed terrific load of network between nodes overall lifetime wireless sensor networks has increased. Furthermore, the proposed algorithm reduced the number of packets sent to the destination during the release [4].

Toloe Honari using genetic algorithm is implemented as a central base station, where the cluster heads are determined so that the network will have minimum energy consumption. In fact, during the course of operations data, selection criteria for the new generation of "minimum energy difference grid" have taken place. Balance and uniformity of the energy consumption of nodes and long life of network is outcome of use genetic algorithms in this study [5].

Dechene and his colleagues have been concerned on how clustering, necessity, benefits, and various combinations of these patterns in wireless sensor networks and effect of clustering in wireless sensor networks, both energy efficiency and optimal use of resources are examined. In addition to the mentioned issues quality of service in wireless sensor networks, a high level of importance. Cases include delay rates, packet loss and network fault tolerance can be achieved by clustering topics. Protocols introduced in this paper meet the above mechanism by mentioned clustering. It should be noted that the optimal number of clusters is a factor that reduces network soldiers, increase efficiency and improve the routing and load distribution in the network. After clustering, in this article some cases need to re-clustering [6].

Handy and his colleagues presented a method for selecting cluster heads in a tree structure with the lowest cost. The protocol based on adaptive clustering or selecting definite cluster heads cause of death postponement of the first node, intermediate node and the last node in the network. That is a smart way to choose cluster heads that select the cluster heads without the knowledge of their location in the network. Some of the ideas in this thesis will be based on the same reference [7].

Mao and his colleagues are considered data collect as a common but critical operation in various applications in wireless sensor networks, using the innovative techniques to improvement the electronic energy consumption and as a result are necessary to long life of the network. Clustering is a useful method for topology control in wireless sensor networks which can increase network scalability and lifetime. In this paper a clustering approach called EECS offered select better cluster heads on accordance with energy consumption loads in the network. Decision-making methods selection cluster heads taking into account the residual energy through local radio communication is effective load balancing among cluster heads. Simulation results show that EECS significantly increasing the network lifetime over 35 % effective [8].

Ray has been applied his original idea on LEACH protocol with the aim of improving the energy consumption and increase the lifetime of the network. In this paper, based on the residual energy of cluster heads nodes, the distance to sink and the nodes in their cluster heads will be periodically decision on future processes. So that cluster heads node with low residual energy, the distance from the sink is high and exist in the network as cluster heads will have any chance to cluster head. Simulation results indicate that the death of the first node, the intermediate node and the last node in the proposed protocol, respectively, 41%, 36% and 25% occur later than LEACH protocol [9].

3. Proposed Protocol

In the proposed protocol, it is assumed that the base station is located in a square in LEACH protocol for data transmission nodes to the base station will consume more energy. This could be one of the causes of the failure are not considered faults in the network, which is not covered the sensor nodes running out their energy and incorrect clustering nodes. This protocol uses fuzzy logical error coverage reduce network lifetime, we will work to increase the network lifetime. At each step of the algorithm based on fuzzy system sensors, each sensor prioritized and the highest priority will be selected to move.

The fuzzy fault tolerant algorithm three criteria for selecting neighbors considered: The criterion measures the amount of overlay neighbor's node with broken sensor, distance neighbor from broken sensor and distance neighbor from the base station. For each neighbor, the values are given as input and the fuzzy system considered priority for each neighbors in comparison to other neighbors for each group, the overall relationship of fuzzy system to choose neighbors would be equation (1).

(1)

$$f(x) = \frac{\sum_{l=1}^{m}(y^{l}\prod_{i=1}^{n}\mu^{l}(x_i))}{\sum_{l=1}^{m}(\prod_{i=1}^{n}\mu^{l}(x_i))}$$

The relation (1), m is the number of fuzzy rules and n is the number of inputs of the fuzzy system, the system will be 3 times. f (x) x is neighbor priority Xth defines each groups. Iy is the answer bet Lth and μ (x_i) is equal to amount of input fuzzy sets. μ (x_i)l given amount of input overlay and distance of each sensor and fuzzy triangular used for each of them.

In the algorithm priority of each neighbor in a group according to its distance from the failed sensor and the overlapping range of the observations will be determined using a fuzzy system. Linguistic variables are used to describe the distance between neighbor and failed sensors, in the short, medium and long are expressed.

The amount of observations overlapping area of each sensor neighbor with his neighbors than the sum of all overlapping neighboring failed sensors sensor is calculated. To obtain this ratio, the total observations overlapping area of all neighboring failed sensor with its neighbors is calculated.

The overlapping ratio each neighbor is equal to the ratio of the observations overlapping area with its neighbors and total observations overlapping area of all the neighboring sensors. So much overlapping neighbors of each sensor are a relative value. Those values said, are normalized value and expressed in the range 0-100, the proportion is amount overlapping neighbor to other neighbors. To express the observations overlapping area of the sensor with their neighbors the linguistic variables very low, low, moderate, high and very high are used. To determine the priority of linguistic variables very bad neighbors, bad, acceptable, good, and very well used. Fuzzy rule system is designed to select neighbors summarized in Table 1.

Table 1. Fuzzy rules for selecting neighbors

Cases	Distance neighboring nodes of faulty sensor	Distance of central station from sensor	Overlapping	Probability of selection
1	High	High	Low	Very low
2	High	High	Middle	Low
3	High	High	High	Relatively low
4	High	Middle	Low	Low
5	High	Middle	Middle	Relatively low
6	High	Middle	High	Middle
7	High	Low	Low	Relatively low
8	High	Low	Middle	Middle
9	High	Low	High	Relatively high
10	Middle	High	Low	Low
11	Middle	High	Middle	Relatively low
12	Middle	High	High	Middle
13	Middle	Middle	Low	Relatively low
14	Middle	Middle	Middle	Middle
15	Middle	Middle	High	Relatively high
16	Middle	Low	Low	Middle
17	Middle	Low	Middle	Relatively high
18	Middle	Low	High	High
19	Low	High	Low	Relatively low
20	Low	High	Middle	Middle
21	Low	High	High	Relatively high
22	Low	Middle	Low	Middle
23	Low	Middle	Middle	Relatively high
24	Low	Middle	High	High
25	Low	Low	Low	Relatively high
26	Low	Low	Middle	High
27	Low	Low	High	Very high

4. Simulation Results

The simulation of the proposed protocol used ns-2 version 2.29 software package to better display the results of the second and third series of simulations have been done. In the first series of nodes in the network 25 nodes, in the second series 50 nodes and in the third series 100 nodes have been compared. To determine the performance the proposed algorithm with the well-known LEACH algorithm has been evaluated. As the graphs will show the proposed algorithm could performs better than other proactive and reactive algorithms.

4.1 Distance of Neighbor to Central Station

It is some nodes based on the location of nodes compare to the central station, using fuzzy variables center is characterized. Central chart is displayed in the fuzzy set. To find the central node, selects the base station of each sensor nodes and average distance nodes from the base station calculates.

4.2 Distance of Neighbor to failed sensor

It is some nodes based on location of node are compare to faulty nodes by fuzzy variables interval are specified. Distance chart is displayed in the fuzzy set. To find the distance select any sensor and the average distance from the failed sensor calculates.

4.3 The amount of overlap

It is some nodes based on location of node than failed sensor is classified, by fuzzy variable the amount of overlapping neighboring sensors than failed sensor in the fuzzy set is shown.

Linguistic variables used to represent overlapping, divided to three level respectively low , medium and high, and three level to represent the central node and the distance to the base station , there are near , intermediate and far away, respectively.

Probability to selects to neighboring sensors are divided motion direction into seven categories: very low, low, relatively low, medium, relatively high, high and very high. Fuzzy base low includes the following rules: If the faulty sensor distance from the nearest neighbor, distance between neighboring sensors near the central station overlapping the neighboring sensors with failed sensor is too high so the probability selection motion direction is very high.

So we'll use $\mathbf{3^3 = 27}$ rules for fuzzy rule base. In order to provide fuzzy sets near, far and medium we use the average of the triangular membership functions and to represent fuzzy sets low, medium, and large, trapezoidal membership functions are used.
Development of Membership functions and related linguistic states is shown in fig.1 and all nodes are compared based on the probability that the sensor is selected motion direction with most likely.

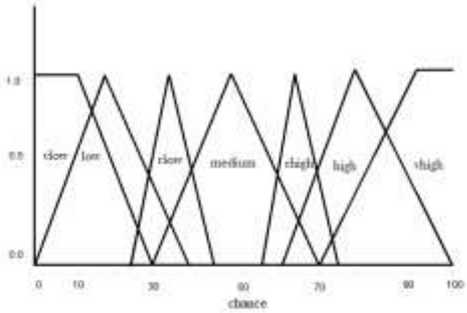


Figure 1. fuzzy sets and probability fuzzy variables

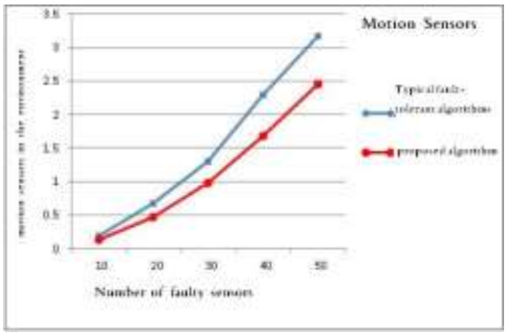


Figure 2. comparing the amount of sensor motion in the network environment

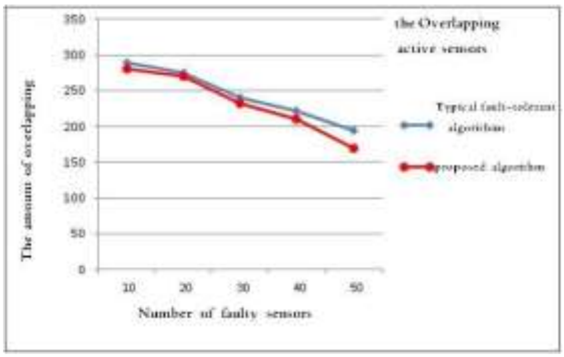
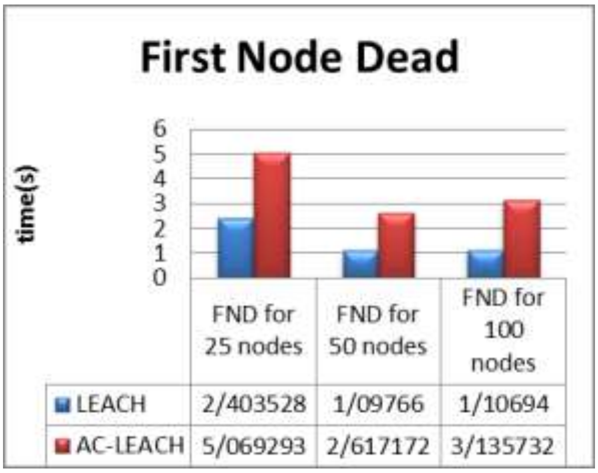


Figure 3. overlapping sensor ratio in the networks

Figures 1 and 2 shown the results of 100 times the performance of the proposed algorithm on network with 100 sensor nodes in a 50 x 50 environment. The results obtained in the above figures is the result of the proposed algorithm both in terms of overlapping amount of sensors in the environment to cover the lost areas and amount of observation overlapping area of sensors. As is clear from the figure, failed sensors how many are reduced down, lost space due to downtime to easily compensate as well as improving network coverage. Simulation results show that the proposed fuzzy algorithm provided better coverage than the conventional fuzzy algorithm and the network's energy consumption is reduced.



	FND for 25 nodes	FND for 50 nodes	FND for 100 nodes
🟦 LEACH	2/403528	1/09766	1/10694
🟥 AC-LEACH	5/069293	2/617172	3/135732

Figure 4. first node dead

As figure 4 is shown; the proposed protocol has better efficiency than LEACH protocol in the parameter of first node death.
As can be seen in 25 nodes that is 5.069293 increasing the number of nodes randomly reduces or increases the duration of the first node death. In total, 100 nodes have shown an increase than in the 50 nodes. But the number has dropped to 50 nodes compared to 25 nodes. The results show a decrease in energy consumption in the network.

5. REFERENCES

[1] Ahmadinia, M., Meybodi, M.R. 2008. "Clustering in Wireless Sensor Networks Using Cellular Learning Automata", Computer Society of Iran.

[2] Akbarzadeh Totonchi, M., Yaghma'ee Moghaddam, M.H 2008. "Clustering self-organizing sensor with movable base stations in wireless sensor networks using fuzzy logic and genetic algorithms", Computer Society of Iran.

[3] Anami, N. 2009. "Reducing energy consumption in wireless sensor networks using neural networks SOM", Dissertation for the degree of Master of Engineering, Computer-oriented software, Information and Communication Technology Department, Payam Noor University Central Tehran Branch.

[4] Roshanzadeh, M. 2009. "Providing an optimal algorithm in terms of energy consumption and reliable sending in the wireless data network", National Conference on Soft Computing and Information Technology, Islamic Azad University, mahshahr.

[5] Toloe Honary, M., Tashtarian. 2007. "Node clustering in the wireless sensor network using genetic algorithm", First Joint Congress on Fuzzy and Intelligent Systems.

[6] Dechene. D. J, Jardali. A. El, Luccini. M and Sauer. A. 2008. ”A Survey of Clustering Algorithms for Wireless Sensor Networks”, Department of Electrical and Computer Engineering the University Of Western Ontario London MS.Thesis, Ontario, Canada.

[7] Handy. M. J, Haase. M, Timmermann. D. 2002. “Low Energy Adaptive Clustering Hierarchy with Deterministic Cluster-Head Selection”, Fourth IEEE Conference on Mobile and Wireless Communications Networks, Stockholm, Erschienen in Proceedings, pp.5.

[8] Mao. Ye, Chengfa. Li, Guihai. C and Jie.Wu. 2005. ”EECS: An Energy Efficient Clustering Scheme in Wireless Sensor Networks”, in Performance, Computing, and Communications Conference, IPCCC 2005., 24th IEEE International, pp. 535-540.

[9] Ray. A. 2012. “Energy efficient cluster head selection in wireless sensor network”, 1st International IEEE Conference in Recent Advances in Information Technology (RAIT), Kolkata, India, pp. 306-311.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 2, 114 - 118, 2014

An Efficient Discovery of High Utility Item Sets from Large Database

Santhamani.V
PPG Institute of Technology
Coimbatore, India

Premkumar.M
Department of CSE
PPG Institute of Technology
Coimbatore, India

Gayathri.A
PPG Institute of Technology
Coimbatore, India

Gokulavani.M
PPG Institute of Technology
Coimbatore, India

Abstract - Identifying frequent items from database and treating each item in a database as equal. However, items are actually differs in many aspects like, profit in real application, such as retail marketing. The difference between items makes a strong impact on the decision making applications, where the values of each items are considered as utilities. Utility mining focuses on identifying the itemsets with high utility like profit, aesthetic value. High utility itemsets mining extends frequent pattern mining to discover itemsets in a large database with utility values above a given threshold. Here we use two algorithms UP-Growth and FP-Growth for mining high utility itemsets and frequent users with a set of effective strategies. The information of high utility itemsets is maintained in a UP-Tree. Candidate itemsets are generated efficiently. Customer Relationship Management (CRM) is incorporated into the system by tracking the customers who are frequent buyers of the different kinds of item sets.

Keywords: Candidate itemsets; Frequent itemset; High utility itemset; Utility mining; data mining.

1. INTRODUCTION

Data mining is the process of showing nontrivial, previously unknown and potentially useful information from large databases. It enables the companies to focus on important information in data warehouses. It can be implemented rapidly on the existing software and hardware platforms and also enhances the value of the existing information resources. Discovering useful patterns hidden in a database plays an essential role in several data mining tasks, such as frequent pattern mining, and high utility pattern mining.

Association rules mining (ARM) is one of the most widely used techniques in data mining and has tremendous applications like business, science etc. Make the decisions about marketing activities such as, e.g., promotional pricing or product placements.

Relative importance of each item in frequent pattern mining is not considered. To address this problem, weighted association rule mining was proposed. In this, weights of items, such as unit profits of items in transaction databases, are considered. In this view, utility mining emerges as an important topic in data mining field. Mining high utility itemsets from the large databases refers to finding the itemsets with high profits. The meaning of itemset utility is interestingness, importance, or profitability of an item to user. Utility of items in a transaction database consists of two aspects:1) the importance of distinct items, and 2) the importance of items in transactions.

A high utility itemset is an itemset if its utility is no less than a user-specified utility threshold; otherwise, it is a low-utility itemset. Mining high utility itemsets from the large databases is not an easy task since downward closure property in frequent itemset mining does not hold. In different way, pruning search space for high utility itemset mining is difficult because a superset of a low-utility itemset may be a high utility itemset. Existing methods often generate a huge set of PHUIs and their mining performance is degraded consequently. The situation becomes worse when database_contain many long transactions or low threshold value are set. The huge amount of PHUIs forms a challenging problem for mining performance since the more PHUIs the algorithm generates, the time consuming process. Major contributions of this work are summarized as follows:

1. Two algorithms, namely Utility Pattern growth (UP-Growth) and FP-Growth, and a compact tree structure, called utility pattern tree (UP-Tree), for discovering high utility itemsets and maintaining important information related to utility patterns within databases are proposed. Efficiently High-utility itemsets can be generated from UP-Tree by two scans of original databases.
2. Several strategies are proposed for facilitating the mining processes of UP-Growth. By maintaining only essential information in UP-Tree, overestimated utilities of candidates can be well reduced by discarding utilities of the items that cannot be high utility.
3. UP-Growth and FP-Growth outperform other algorithms substantially in terms of execution time, especially when database contain lots of long transactions or low minimum utility thresholds are set.

1.1 Preliminary

Given a finite set of items L = {l_1,l_2,….l_n}, each item l_p (1≤ p ≤n)has a unit profit pr(l_p). An itemset X is a set of k distinct items {l_1, l_2 ,….l_k}, where l_j £ L,1≤j≤k.k is the length of X. An itemset with length k is called a k-itemset. A transaction database D = {T_1, T_2,…,T_m} contains a set of transactions, and each transaction Td(1 ≤ d ≤m) has a unique identifier d, called TId. Each item l_p in transaction Td is associated with a quantity q (l_p,T_d), that is, the purchased quantity of l_p in T_d.

TABLE 1

An example Database

TID	Transaction	TU
T1	(A,2)(B,3)	12
T2	(B,2)(C,2)(D,1)	15
T3	(C,1)(D,2)	7
T4	(A,1)(B,1)(C,3)	20

TABLE 2

Profit Table

Item	A	B	C	D
Profit	3	2	5	1

Definition 1

An itemset is no less than user- specified minimum utility threshold which is called high utility itemset, Otherwise it low-utility itemset.

For example, in Tables 1 and 2,

u({A},T1) = 3×2=6;

u({AB},T_1) = u({A},T_1)+ u({B},T_1)
=6+6 = 12;

u({AB})= u({AB},T1)+u({AB},T4)
=12+5 =17;

If min_util is set to 17, {AB} is a high utility itemset.

Definition 2

Transaction-weighted utility(TWU) of an itemset X is the sum of the transaction utilities of all the transaction containing X,wich is denoted as TWU(X).

Property 1: (Transaction-weighted downward closure)

Any subset of a high transaction-weighted utilization itemset must also be high in transaction-weighted utilization it is called transaction weighted downward closure (TWDC).

Downward closure property can be maintained in utility mining by applying the transaction weighted utility. For example, TU(T1) = u({ABC},T1)=17; TWU ({A})=

TU(T1)+TU(T4)=17+28= 45; If min_util is set to 30, {A} is a HTWUI.

The challenge of utility mining is restricting the size of the candidate set and simplifying the computation for calculating the utility.

Problem statement- The problem is producing a large number of candidate itemsets for high utility itemsets. Apriori based algorithms prune candidate itemsets, however algorithms need to test all candidates. Moreover, they must repeatedly scan a large amount of the original database in order to check if a candidate item is frequent or not. It is inefficient and ineffective.

2. RELATED WORK

One of the well-known algorithms for mining association rules is Apriori [1], which is used for efficiently mining association rules from large database. Pattern growth-based association rule mining algorithms [3], [5] such as FP-Growth [3] were afterward proposed. It achieves a better performance than Apriori-based algorithms since it finds frequent itemsets without generating any candidate itemset and scans database just twice.

In a frequent itemset mining, the importance of items to users is not considered. So, the topic called weighted association rule mining was introduced. The weighted association rule mining (WARM) considers the importance of items, in some applications such as transaction databases, items' quantities in transactions are not_yet consider. So, the issue of high utility itemset mining is raised. And many studies [2], [4], [6], [7], [8] have addressed this problem.

Two phase algorithm has been proposed [4] which is composed of two mining phases. In phase 1, an Apriori-based level wise method is used and the complete set of HTWUIs is collected. In phase 2, an additional database scan is computed for identify HTWUIs. The two phase algorithm also reduces search space by using TWDC property, but it still produces too many candidate items to obtain HTWUIs and requires multiple database scans. To overcome this, an isolated items discarding strategy (IIDS) was introduced to reduce the number of candidate items. This algorithm uses a candidate generation-and-test scheme for finding high utility items but it still scans database for several times.

To avoid scanning database too many times and to generate HTWUIs, a tree based algorithm has been proposed, namely IHUP. In IHUP-Tree the information about itemsets and their utilities are maintained. This algorithm has three steps: 1) Create IHUP-Tree, consists of an item name, TWU value and support count, 2) HTWUIs are generated with the help of FP-Growth, 3) The original database has been scanned once, in which High utility itemsets are identified. From the above steps, there are many HTWUIs, thus the performance of an algorithm became a critical issue.

Hence the overestimated utilities of itemsets have to be reduced by applying proposed method.

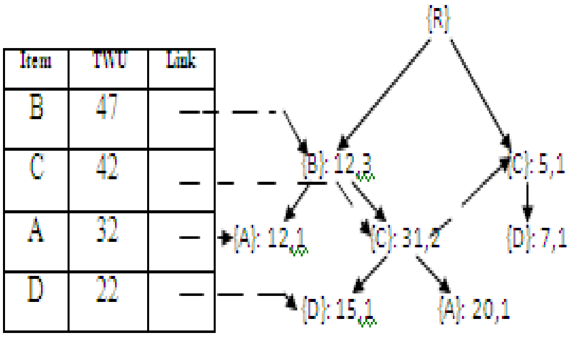
3. PROPOSED METHODS

The framework of proposed methods consists of four steps: 1) Scan the database twice to construct a UP-Tree; 2) Generate PHUIs (Potential High Utility Itemsets) from UP-Tree by using UP-Growth; 3) Identify actual high utility items from the set of PHUIs; 4) Identify frequent users by using FP-Growth and incorporating CRM;

3.1 Data structure: UP-Tree

To improve the mining performance and avoid repeated scanning of original database, we use a UP-Tree structure. It is used to maintain the information of transactions and high utility itemsets.

In a UP-Tree each node N consists of N.name (node's item name), N.count node's support count), N.nu (node's node utility, i.e., overestimated utility of the node), N.parent (records the parent node of N), N.hlink (node link points to a node which name is the same as N.name) and a set of child nodes.



Item	TWU	Link
B	47	
C	42	
A	32	
D	22	

Figure 1. UP-Tree

3.2 Mining Method: UP-Growth

After constructing a UP-Tree, a basic method for generating PHUIs is to mine UP-Tree by UP-Growth by pushing two more strategies into the framework of FP-Growth. From the strategies, overestimated utilities of itemsets can be decreased and thus the number of PHUIs can be further reduced.

3.3 Identify High Utility Itemsets

After finding all PHUIs, the third step is to identify high utility itemsets and their utilities from the set of PHUIs by scanning original database once. This step is called phase II. However, in previous studies, two problems in this phase occur: 1) number of HTWUIs is too large; and (2) scanning original database is very time consuming. In our work, overestimated utilities of PHUIs are smaller than or equal to TWUs of HTWUIs since they are reduced by the proposed strategies. So, the number of PHUIs is much smaller than that of HTWUIs. Hence, in phase II, our method is much efficient than the previous methods._Although our methods generate fewer candidates in phase I, scanning original database is still time consuming since the original database is large and it contains lots of unpromising items. In our framework, high utility itemsets can be identified by scanning reorganized transactions. Since there is no unpromising item in the reorganized transactions, I/O cost and execution time for phase II can be further minimized. This technique works well especially when the original database contains lots of unpromising items.

3.4 Identify frequent users and CRM

The item sets that are both high frequent and high utility can be obtained using FP-Growth. From the basic framework of this algorithm the different kinds of item sets namely high utility high frequent, high utility low frequent, low utility high frequent and low utility low frequent are generated. Then Customer Relationship Management (CRM) is incorporated into the system by tracking the customers who are frequent buyers of the different kinds of item sets.

4. CONCLUSION

Generate potentially high utility itemsets using utility pattern with two database scans. This combines with the frequent pattern to provide better performance and gives best solution for time consumption. Apriori algorithm requires multiple time databases scanning. To find long patterns it may need too many database scanning that is quite time consuming. Meantime, while processing data sets that contain long patterns, it generates too many candidates and subsequences of frequent patterns. To solve these problems we are using high utility pattern, which avoids the costly candidate generation and requires only two times database scanning. The first pass finds all frequent items, and the second pass constructs compact data structure using the high utility items which are used for storing compressed, crucial information about high utility patterns.

Customer Relationship Management (CRM) is incorporated into the system by tracking the customers who are frequent buyers of the different kinds of item sets.

5. REFERENCES

[1] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” Proc. 20th Int’l Conf. Very Large Data Bases(VLDB), pp. 487-499, 1994.

[2] A. Erwin, R.P. Gopalan, and N.R. Achuthan, “Efficient Mining of High Utility Itemsets from Large Data Sets,” Proc. 12th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 554-561, 2008.

[3] J. Han, J. Pei, and Y. Yin, “Mining Frequent Patterns without Candidate Generation,” Proc. ACM-SIGMOD Int’l Conf. Management of Data, pp. 1-12, 2000.

[4] Y. Liu, W. Liao, and A. Choudhary, “A Fast High Utility Itemsets Mining Algorithm,” Proc. Utility-Based Data Mining Workshop, 2005.

[5] J. Pei, J. Han, H. Lu, S. Nishio, S. Tang, and D. Yang, “H-Mine: Fast and Space-Preserving Frequent Pattern Mining in Large Databases,” IIE Trans. Inst. of Industrial Engineers, vol. 39, no. 6,pp. 593-605, June 2007.

[6] B.-E. Shie, V.S. Tseng, and P.S. Yu, “Online Mining of Temporal Maximal Utility Itemsets from Data Streams,” Proc. 25th Ann. ACM Symp. Applied Computing, Mar. 2010.

[7] V.S. Tseng, C.J. Chu, and T. Liang, “Efficient Mining of Temporal High Utility Itemsets from Data Streams,” Proc. ACM KDD Workshop Utility-Based Data Mining Workshop (UBDM ’06), Aug. 2006.

[8] V.S. Tseng, C.-W. Wu, B.-E. Shie, and P.S. Yu, “UP-Growth: An Efficient lgorithm for High Utility Itemsets Mining,” Proc. 16th ACM SIGKDD Conf. Knowledge iscovery and Data Mining (KDD ’10), pp. 253-262, 2010.

[9] C.F. Ahmed, S.K. Tanbeer, B.-S. Jeong, and Y.-K. Lee, “Efficient Tree Structures for High Utility Pattern Mining in Incremental Databases,” IEEE Trans. Knowledge and Data Eng., vol. 21, no. 12,pp. 1708-1721,Dec.2009.

[10] C. Creighton and S. Hanash, “Mining Gene Expression Databases for Association Rules,” Bioinformatics, vol. 19,no.1,pp.79-86,2003.

[11] M.Y. Eltabakh, M. Ouzzani, M.A. Khalil, W.G. Aref, and A.K.Elmagarmid, “Incremental Mining for Frequent Patterns in Evolving Time Series Databases,” Technical Report CSD TR#08-02, Purdue Univ., 2008.

[12] C.H. Lin, D.Y. Chiu, Y.H. Wu, and A.L.P. Chen, “Mining Frequent Itemsets from Data Streams with a Time-Sensitive Sliding Window,” Proc. SIAM Int’l Conf. Data Mining (SDM ’05), 2005.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 2, 119 - 124, 2014

Swarm Intelligence Based Optimization for Web Usage Mining in Recommender System

Manisha Sajwan
DIT University
Dehradun, India

Kritika Acharya
DIT University
Dehradun, India

Sanjay Bhargava
DIT University
Dehradun, India

Abstract: Nowadays, the web has become one of the most effective and efficient platform for information change and retrieval .Due to heterogeneity and unstructured nature of the data available on the WWW, web mining uses various data mining techniques to discover useful knowledge from web hyperlinks, page content and usage log. This research introduces the theoretical foundations of Swarm Intelligence and Design, implementation of swarm optimization algorithm. The Swarm Intelligence optimization and data mining technique can be used together to form a method which often leads to the result. Design and implementation of a web mining system based on multi-agents technology will reduce the information overload and search depth. This is helpful to users using the web within a platform for e-commerce or e-learning.Swarm Intelligence is an efficient technology that deals with natural and artificial system. It provides an efficient way for finding optimal solution. During the past few decades researches are trying to use these techniques to solve many problems in various fields. Recommender System is the one of the most important application of e-commerce and it plays vital role in understanding the user's behaviour or interest by which it increases the profit of sales or usage of services of website. This paper describes a swarm intelligence optimization for web mining to find the optimal solution and based on that process is done.

Keywords: Swarm Intelligence Optimization, Natural Inspired Technique, Web Mining, Recommender System, E- Commerce, Target Marketing.

1. INTRODUCTION

Swarm Intelligence is the collective behaviour of natural and artificial system. The concept is employed by Gerado, Beni and Jing Warm in 1989 in the context of cellular robotics system. Natural Inspired Computing Techniques such as swarm intelligence has the ability to solve many combinatorial problems and provide optimal solution. Swarm Intelligence deals with natural and artificial system that provides an effective and an efficient way for finding optimal solution. The combinational environment of natural and artificial system is called decentralized system. In this type of environment user has difficult to find the optimal solution so, this can be overcome by recommendation process. By using this process we can find nearest neighbour with users of similar likes and dislikes and there for we can generate an effective recommendation.

Data Mining and Swarm Intelligence may seem that they do not have many properties in common. However the combinational approach of data mining and swarm intelligence can be used to generate optimal result, even other methods or approaches would be too expensive or difficult to implement.

This paper focuses on two parts. First part describes Swarm Intelligence Based Web Usage Mining which describes a study on collective behaviour of ant, bees etc, mining techniques and Knowledge Representation and Discovery. Second part describes the user profile are selected based on neighbourhood usage of Swarm Intelligence.

This paper is organised as follows:
Section 2 outlines a deep study on swarm intelligence and knowledge representation and discovery. Section 3 outlines the overview of Swarm intelligence algorithms. Section 4 describes web usage mining. Section 5 describes Recommender System, E-Commerce and Target Marketing. Section 6 outlines Web Page Recommendation using Swarm Intelligence.

2. SWARM INTELLIGENCE, KNOWLEDGE REPRESENTATION AND DISCOVERY

Swarm behaviour can be seen in bird flocks, fish schools, as well as in insects like mosquitoes and midges. Swarm is an aggregate with cohesion, but a low level polarization (parallel alignment) among members. Swarm Intelligence (SI) is an innovative distributed intelligent paradigm for solving optimization problems that originally took its inspiration from the biological examples by swarming, flocking and herding phenomenon in vibrates.[7],[13]

Main principles of collective behaviour-

Homogeneity- Every bird in flock has the same behaviour model. The flock moves without a leader, even though temporary leaders seem to appear.

Locality- The motion of each bird is only influenced by its nearest flock mates. Vision is considered to be the most important senses for clock organization.

Collision Avoidance- Avoid with nearby flock mates.

Velocity Matching- Attempt to match velocity with nearby flock mates.

Flock Centring- Attempt to stay close to nearby flock mates.

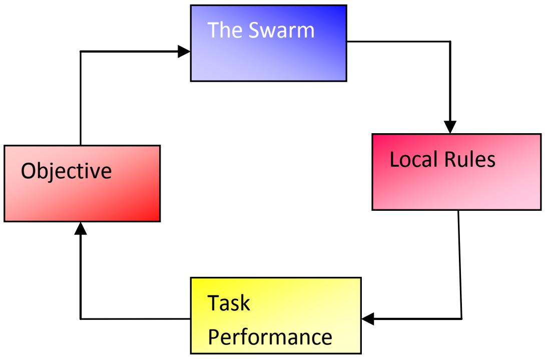


Figure 1. Simple Scheme of a Swarm

Since 1990, several collective behaviour inspired algorithms have been proposed. The application areas of those algorithms refer to well studied optimization problem like NP-hard problems(Travelling Salesman Problem, Quadratic Assignment Problems, Graph Problems), Network Routing, Clustering, Data Mining, Job Scheduling, etc.

2.1 Knowledge Representation

Knowledge Representation is necessary to solve a real word problem using large amount of data by manipulating that knowledge.

Before going through Knowledge Representation And Discovery, we have to focus on two entities.

1-Fact (Universal truth)

2-Representation of fact in some chosen form

Predicate Logic is the best method to represent knowledge. To use predicate in our application, we can implement an interface predicate in our programming like in java. This interface provides the framework for filtering of database.

2.2 Knowledge Discovery:

The following model as shown in figure 2 & 3 presents how to discover knowledge from the large set of information. The knowledge discovery process seeks new knowledge in same application domain. It is defined as a non-trivial process of identification of valid novel potentially useful and ultimately understandable patterns in data. It consist of many step each attempting to complete a particular discovery task and each accomplished by the application of discovery method.

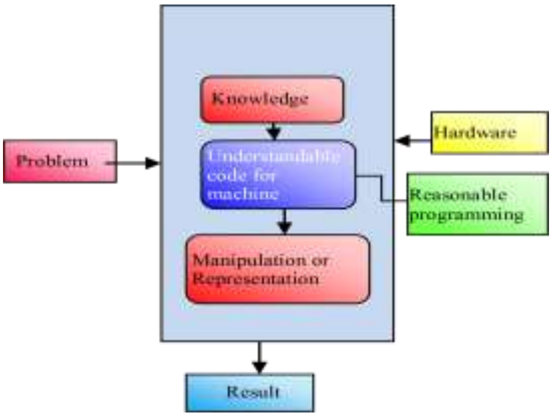

Knowledge
Problem
Understandable code for machine
Hardware
Reasonable programming
Manipulation or Representation
Result

Figure 2. Knowledge Representation

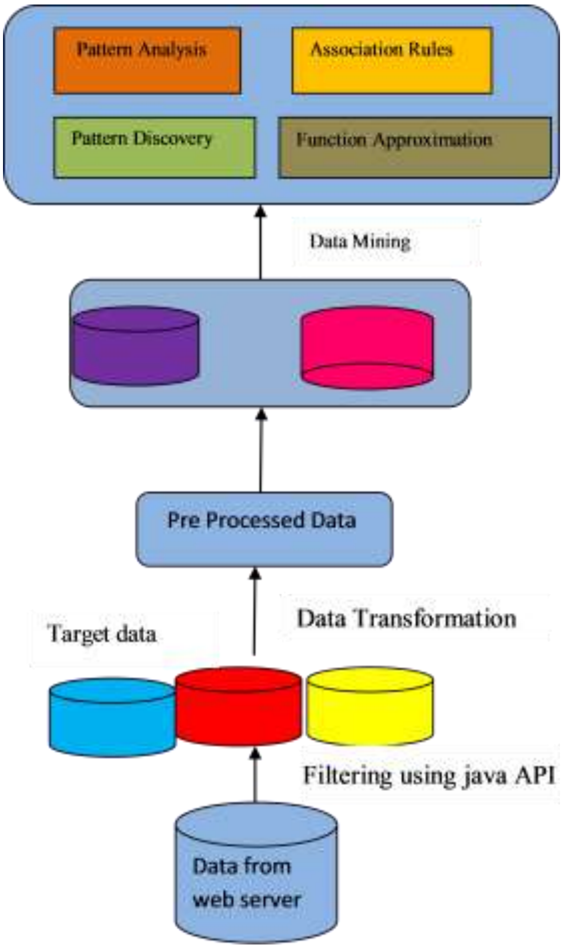


Figure 3. Knowledge Discovery

3. SWARM INTELLIGENCE ALGORITHMS:

Several collective behaviours such as bees, ant, firefly, cuckoo, etc., inspired algorithm have been proposed. These algorithm provide problem solving ability and study on this makes human understandable in nature. The main principles of collective behaviour of swarm are homogeneity, locality, collision avoidance, velocity matching and flock clustering. Some of the popular SI algorithms are as follows:

3.1 Ant colony optimization

Ant colony optimization (ACO) is a behaviour of ants that finds its path between food source and colony. Ant while returning colony lay down pheromone, which directs the search of future ants on the same path. This helps to find the optimal solutions. They use the environment as a medium of communication. Some of popular variations of ant colony are edges, Max Min ant system, convergence and pheromone update. ACO has been useful to solve many combinatorial problems and optimum solutions it is also used in web usage mining to each user. ACO deals with artificial systems that inspired from the foraging behaviour of real ants, which are used to solve discrete optimization problem. The main idea is the indirect communication between the ants by means of chemical pheromone trials, which enables them to find short paths between their nest and food. The study of ant colonies behaviour and their organizing capabilities is of interest to knowledge retrieval or management and decision support systems sciences, because it provides models of distributed adaptive organization, which are useful to solve difficult classification, clustering and distributed control problems.[1],[11],[12]

3.2 Bat algorithm

Bat algorithm (BA) is a Meta heuristic optimization inspired by the echolocation of micro bats. Bats are one kind of mammals that have the capacity of echolocation. These bat produce loud sound based on the echo reached it identifies the obstacles and severs even in dark. Many researchers found the behaviour of bat has the solution to many complex problems.

3.3 Cuckoo search

Cuckoo search (CS) is swarm intelligence inspired by the some cuckoo species they lay their eggs in the nest of another birds. The breeding behaviour is applied to many optimization problems. Each egg represents the solution. Recent studies suggest that comparison between CS searches with PSO says CS is robust results.[2]

3.4 Firefly algorithm-

Firefly algorithm (FA) inspired by the flashing behaviour of fireflies. Firefly algorithm consists of three rules:

1. No firefly will be attracted to another since they are unisex.

2. Attractiveness is proportional to brightness that is less brighter one is attached to brighter one.

3. If no firefly is brighter than the giver firefly then they are moved randomly.

In fact the variants of firefly algorithm are discrete firefly algorithm, multi objective FA and so on. It is applied on image processing, clustering, continuous optimization, etc.

3.5 Particle swarm optimization-

Particle swarm optimization (PSO) is a swarm intelligence global optimization technique. It was founded in1995 by James Kennedy and Russell Eberhart to model the convergence behaviour of a flock of birds. PSO is a population based algorithm and is initialized with a population of random solutions, called particles. Each particle is associated with a velocity. Particles fly through the search space with velocities which are dynamically adjusted according to their historical behaviours. Therefore, the particles have the tendency to fly towards the better and better search area over the course of search process. It is mainly inspired by the social behaviour of bird flock and fish school. In this if one particle identifies path for food source or protection then rest of swarm follows it automatically even if it is in opposite direction. The birds also have the capacity to smell the food so it finds the optimum solution to find the food. Using the pre-defined fitness function the performance of each particle is measured.[3], [9],[14]

4. WEB USAGE MINING

Web Usage Mining aims to capture users interacting with the web. Web Usage Mining is the non-trivial process to discover valid, novel, potentially useful knowledge from web data using the data mining techniques. Data stored in usage logs can be used for solving navigational problems, improving web search, recommending queries, suggesting web-sites and enhancing performance of search engines.

Web Mining is the non-trivial process to discover valid, novel potentially useful knowledge from web data using the data mining technique. It may give information that is useful for improving the services offered by web portals. Web mining considered as the most efficient net tool in converting the meaningless information into meaningful in the internet environment , processing the data, extracting the data, and making route map b y the data acquired web 3.0.[6]

Web Usage Mining is the technique or application of data mining to the data generated by the interaction of user with the web servers, this kind of data stored in server logs, represents a valuable source of information, which can be exploited to optimize the document retrieval task or to better understand and thus satisfy user’s need.

For recommendation of web page, here present a graph based approach, leverages the user browsing logs to identify early adaptors, there user discover interesting content before author and monitoring their activity we can find web pages to recommend.[10]

Usually web usage data is collected from web server which is in form of log file. The mined knowledge from log file is used by the companies to establish better customer relationship by giving them exactly what they need. The companies can find attract and retain customers which help them to improve their business performance.

5. RECOMMENDER SYSTEM

An information filtering technology, company used on e-commerce website that uses a collaborative filtering to present information on items and products that are likely to be of interest to the reader. In presenting the recommendations, the recommender system will use detail of the registered user’s profile and opinions and habits of their whole community of users and compare the information to reference characteristics to present the recommendations. Recommendation helps to group similar type of users together. Recommendation gives suggestions about user based on their information previously provides or currently currently visited web page.

An example of recommender System is WhatShouldireadnext.com. A site where users can enter a title of a recent book they have read and enjoyed to see recommended books that they are likely to also enjoy.

It is often necessary to make choices without sufficient personal experience of the alternatives. In everyday life, we rely on recommendations from other people either by word of mouth, recommendation letters, movies, book reviews printed in newspapers, or general surveys. In a Typical Recommender System people provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients. The system’s value lies in its ability to make good matches between the recommender and those seeking recommendations.

The developers of the first Recommender System, Tapestry coined the phrase “Collaborative Filtering”.

The good examples of Recommender Systems are-

1. Offering news articles to on-line newspaper readers, based on a prediction of reader interests.
2. Offering customers of an on-line retailer suggestion about what they might like to buy based on their past history of purchases and for product searches.

Recommender Systems allow to learn user preferences and to make recommendations. They can be employed to recommend products (e.g. news, photos, movies, music, etc).

Recommender System can be: (i) Content based, the system recommend items similar to the ones the user preferred in past. (ii) Collaborative-Filtering based, the system recommends items that people with similar tastes liked in the past and (iii) Hybrid, the system combines content and collaborative-filtering based methods.

Target Marketing- Main goal of Target Marketing is to identify group of users or customers with similar behaviour so that one can predict the customer's interest and make proper recommendations to improve.

6. WEB - PAGE RECOMMENDATION USING SWARM INTELLIGENCE:

The systems for web page recommendation are based on collaborative filtering approaches. The idea is to exploit the user browsing logs (that is click information), in order to identify users with similar likes and dislikes. These recommender systems deals with very high levels of noise, since visiting a web page is not a clear indication of interest as renting a DVD. Search engine help people to search for information on the internet, but the web search is effective only when the users have a clear idea of what they want. Often, people have no specific information need. Recommender system produce suggestions and they are effective in static and relatively noise-free environment. The design of recommender system for web content poses significant challenge due the dynamic nature of webpage, and the high level of noise introduced by the analysis of the user-browsing data.[4]

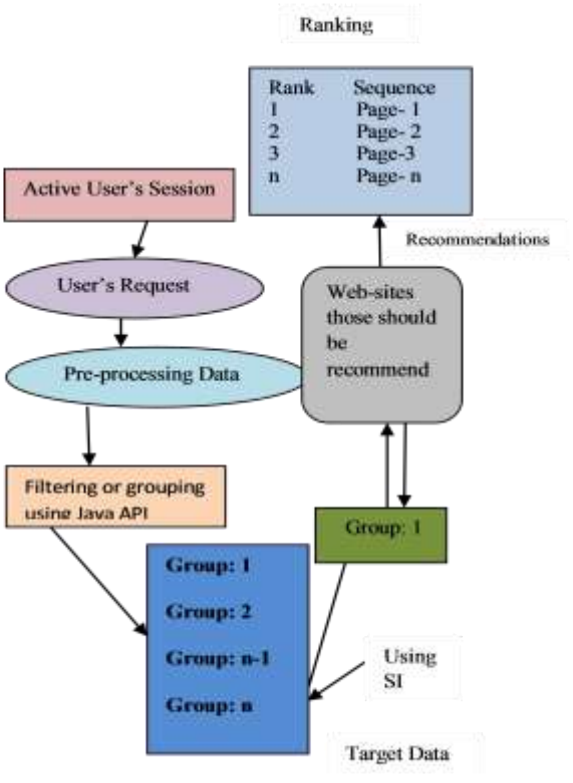


Figure 4. Block Diagram to understand web recommendations

For web page recommendation, here we will use a graph based approach. This algorithm is proposed to use this graph to identify those users who discover interesting pages before this. By tracking the browsing activity of early users we can identify new interesting pages early and recommend these pages to users who should interest with early users.

The optimal solution on diversity is found by recommendations. Based on customer's likes and dislikes they are usage profile is grouped. To generate the recommendations for active user [5], we can follow the following steps:

6.1 Approach

1. User Profile Management
2. Identifying the Optimal Nearest Neighbour Profile
3. Profile Matching

6.1.1 User Profile Management

The Very First step is to generate or collect usage profile of each user from the given website. The profile [U, J] represents the usage profile of user U for J item. The early–user graph G is an attributed, directed, and weighted graph, where nodes correspond to user and an edge between two users, u and v, express the fact that the two different users visited the same site. We generate a graph using tuple (u, p, t, z). Where: (i) u is the identifier of the user. (ii) p is the URL of page visited by user. (iii) t is the timestamp of the visit page by u at p. (iv) z is the optional attribute, we can use this attribute to hold coordinates of topic that has been liked by user at any particular page.

Let, there be n users, $u_1,u_2,u_3......u_{n-1},u_n$ and p_j is the page that has been visited by all users at different timestamp values that is $t_1,t_2,t_3......t_{n-1},t_n$. So the chronologic access will be:- $V[p_j]$: $[(u_1, t_{1j}), (u_2, t_{2j}), (u_3, t_{3j}), ..., (u_{n-1}, t_{n-1j}), (u_n, t_{nj})]$. Where t_{ij} is the timestamp value of the first visit of user i to the page j, and z denotes the co-ordinates of the topic on particular page that has been liked by users.

6.1.2 Identifying the Optimal nearest neighbour Profile

The neighbourhood selection is processed to find the optimal solution the users that are most nearest user to current active user. The best nearest similarity is found nearest neighbourhood selection. The optimal nearest user profile is identified using the Swarm intelligence Optimization Techniques. By calculating relative position, time distance function, edge-weight for the given active user is selected. The Swarm intelligence can be ACO, PSO, Bat, Cuckoo Search, and Firefly.

6.1.3 Profile Matching

Let R(u) be the relative position of user computed by considering the position of user in the access list. The Profile Matching is used to calculate distance between two different profiles using time distance function. We can treat the relative position and timestamp value as the coordinates of particular user like $(u_{i,tij})$, where u denotes the user and t denotes the timestamp value of user for page p that has been liked by user u. The Time distance can be defined as:

Time Distance $[U,J] = \sqrt{[(u_{n-1-}u_n)^2 + (t_{(n-1)j}-t_{nj})^2]}$

Using this time distance function, we can calculate similarities between user profiles and the nearest optimum is chosen. The Recommender System can easily choose the profiles whose Time Distance Function is above a certain threshold value v_{th} as the neighbourhood of u.

Weight of edge from user u to v, i.e. w[u,v], represents that a page visited by u is then visited by v. The edge weight is defined by Time Distance between two different profiles.

The nearest Neighbourhood Selection process is used to find the optimal solution, the users that are most nearest to current user. Let us assume that U is the current working user, the users with similar likes are treated as the optimum solution for user U. The profiles are selected from the database. Based on threshold value profiles are grouped together. The usage profiles are best similar among them will be selected.

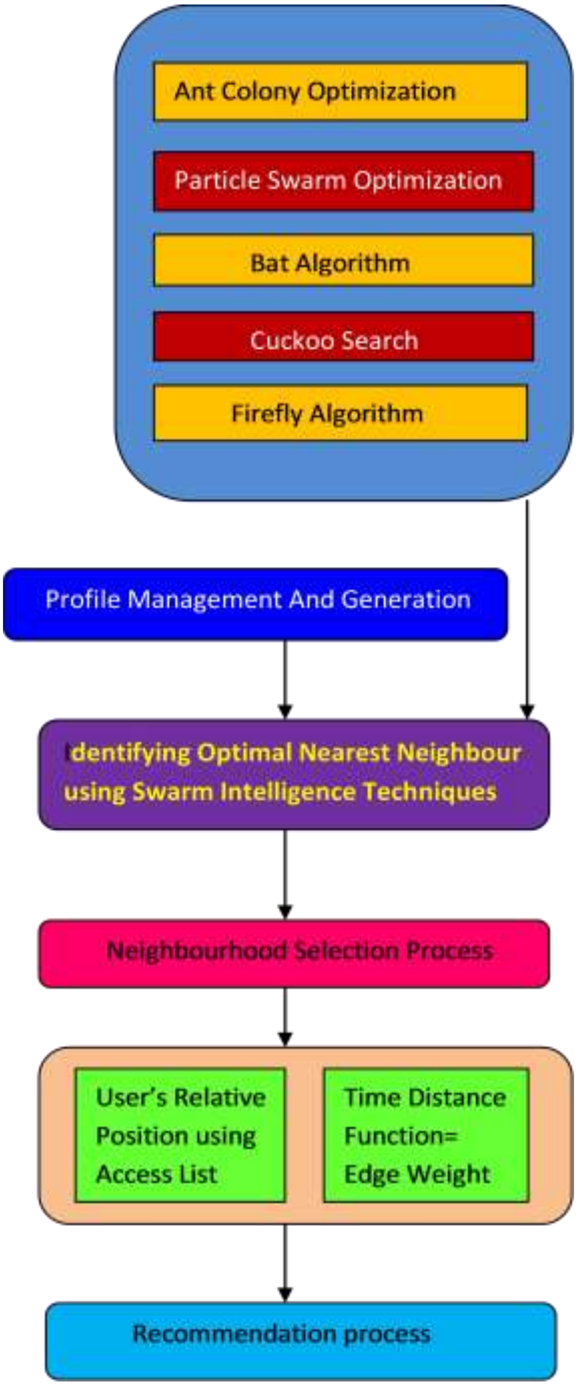

Ant Colony Optimization
Particle Swarm Optimization
Bat Algorithm
Cuckoo Search
Firefly Algorithm
Profile Management And Generation
Identifying Optimal Nearest Neighbour using Swarm Intelligence Techniques
Neighbourhood Selection Process
User's Relative Position using Access List
Time Distance Function= Edge Weight
Recommendation process

Figure 5. Web-Page recommendation Using Swarm intelligence Optimization

Web personalization implicitly or explicitly collect the data from the user. The above figure represents the framework the way recommendation is obtained. After the profile is generated obviously, it is ready for recommendation. The next step after profile generation and management is the nearest neighbourhood selection it can be processed by any one of five popular Swarm Intelligence Techniques. Next step to calculate the Relative position of user and Time Distance Function based on the threshold value. This relative position is treated as the weight between edges. The final stage is the recommendation process is generated. To improve the relevance of our recommendations, we rank recommendations by using the relative position of user u from whom the recommendations originates, as well as the edge weight w[u,v]. Additionally, we can use the page topic to improve or boost scores of page whose topics match the interests of user u. When a page is suggested by many early active users, the final recommendation from u to v is the sum of contributions of the early active users for v.

7. CONCLUSION

The key contribution of this paper is web usage mining using Swarm Intelligence Optimization. This work has been shown that recommendation process is carried out with Swarm Optimization Techniques. This research introduces the theoretical foundations of the biological motivation and Swarm Intelligence Optimization Techniques with a focus on Web-Page Recommendation. The interpretation of result can be used in focalized marketing strategy like direct marketing and target marketing. Concerning the early-adopter graph, we can plan to use different models to learn the edge-weights. Then we can investigate the applications of early-adopter model to other domains.

8. REFERENCES

[1] Abdurrahman et. al., Classification of web user in web usage mining using ant colony optimization algorithm, Doctor Dissertation, Institute of technology, 2009.

[2] Ang X. S. and Deb S. (2010a) Engineering Optimization by Cuckoo Search, Int. J. Math. Modelling &Num. Optimization, Vol. 1, 330-343.

[3] J. Kennedy, R. Eberhart, 1995. Particle Swarm Optimization, IEEE International Conference on Neural Network, Vol. 4, pp. 1942-1948.

[4] N Golovin, E. Rahm: Reinforcement learning Architecture for Web Recommendation, Proceedings on Information System.

[5] I. Mele., F. Bonchi, and A. Gionis. The early-adopter graph and its application to web-page recommendation. In CIKM, 2012.

[6] J. Srivastava , R. Cooley, M. Deshpande, and P.-N. Tan. Web Usage Mining: discovery and application of usage patterns from web data. SIGKDD Explor . Newsl. , 1(2):12-23, 2000.

[7] A. Abraham, He Guo, and Hongo Liu, Swarm Intelligence: Foundations, Perspective and Applications. Studies in Computational Intelligence (SCI), vol.26,pp.3-25.2006.

[8] A. Khosla, Shakti Kumar, K.K. Aggarwal, and Jagatpreet Singh. A Matlab Implementation of Swarm Intelligence based Methodologies for Identification of Optimized fuzzy Models, Studies in Computational Intelligence (SCI) vol.26, pp. 175-

184, 2006.(this reference is taken by me only for study point of view)

[9] J. Kennedy, R. Eberhart, 1995. Particle Swarm Optimization, IEEE International Conference on Neural Network, Vol. 4, pp. 1942-1948.

[10] Abdurrahman et al., Classification of web user in web usage mining for analyzing unique behaviour of web user, International Conference on Electrical Engineering and Informatics, 2007, pp. 356- 359.

[11] Abraham A, Ramos V (2003) Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming, 2003 IEEE Congress on Evolutionary Computation (CEC2003), Australia, IEEE Press, ISBN 0780378040, 1384-1391.

[12] Dorigo M, Blum C (2005) Ant colony optimization theory: A survey. Theoretical Computer Science, 344(2-3), 243-278.

[13] Liu Y, Passino KM (2000) Swarm Intelligence: Literature Overview, http://www.ece.osu.edu/ passino/swarms.pdf.

[14] Pomeroy P (2003) An Introduction to Particle Swarm Optimization,
http://www.adaptiveview.com/articles/ipsop1.html.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 2, 125 - 129, 2014

Inferring User Goals Using Customer Feedback and Analyzing Customer Behavior

Gayathri A.
PPG Institute of Technology
Coimbatore, TamilNadu, India

Nandhakumar C.
PPG Institute of Technology
Coimbatore, TamilNadu, India

Gokulavani M.
PPG Institute of Technology
Coimbatore, TamilNadu, India

Santhamani V.
PPG Institute of Technology
Coimbatore, TamilNadu, India

Abstract: Many enterprises devote a significant portion of their budget to new product development (NPD) and marketing to make their products distinctive from those of competitors, and to know better the needs and expectations of consumers. Hence, knowledge and suggestions on customer demand and consumption experience has become an important information and asset for enterprises. Inferring user search goals are very important in improving the efficiency. For this, feedbacks are obtained from the customer. The submitted feedbacks are clustered as feedback sessions. Pseudo-documents are generated to better understand the clustered feedbacks. K-means clustering algorithm is used to cluster the feedbacks. These feedbacks are very useful in development of new product. Ranking model is used to provide ranks to the products based on the customer feedbacks. Hence knowledge and feedback from customers has become important information. Product design is integrated with the knowledge of customers. Users may also pose their questions about the products which are added when it is suitable. Hence customer behaviour can be analysed from their posed questions and response. Finally, evaluation criterion is described to evaluate the performance of new product.

Keywords: feedback sessions, k-means, pseudo-documents, customer behaviour, ranking model

1. INTRODUCTION

Nowadays data mining has attracted a great deal of attention in the information industry and in society as a whole, due to the wide availability of large amounts of data and the imminent need for turning such data into useful knowledge and information. The information and knowledge gained can be used for many applications ranging from market analysis, customer retention, fraud detection, to production control and science exploration.

Clustering is the most important concept used here. Clustering analyzes data objects without consulting a known class label. The objects are grouped or clustered based on the principle of maximizing the intra class similarity and minimizing the inter class similarity.

Knowledge of the customers and the product itself reflect the needs of the market. Product design and planning for production lines be integrated with the knowledge of customers and market channels. The knowledge of customers and market channels is transformed into knowledge assets of the enterprises during the stage of NPD. The priori algorithm in data mining is a methodology of association rule, which is implemented for mining demand chain knowledge from channels and customers. Knowledge extraction is illustrated as knowledge patterns and rules in order to propose suggestions and solutions to the case firm for NPD and marketing. K-means clustering algorithm is a method of vector quantization originally from signal processing, that is mainly for cluster analysis in data mining. It aims to partition *n* observations into k clusters in which each observation belongs to the cluster with the nearest centre, serving as a prototype of the cluster.

In this paper, customers have the privileges to register their personal details and also have the privileges to create their own user name and password. Then Customer can login to enter which kind of product they want exactly and complaints about their product which they are used. Customers can provide the suggestions and ratings about the product. To get suggestion from every user individually is a difficult task for a manufacture company. Therefore from the customer feedbacks, feedback sessions are proposed. Then, we propose a method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. Ranking model is used to provide the ranking for the products. By providing the ranking for the product the dealers can clearly understand the betterment of products and can easily find out the frequently used products. This can be very useful in improving the product in the new product development stage. Apriori algorithm is a methodology of association rule of data mining, is used to find out the frequently used products. In addition to this a method is provided to analyse the customer behaviour. Here users can pose their questions, in which answers can give by other customers. From their responses we can predict the user expectations and needs. Since the evaluation of clustering is also an important problem, evaluation is described to evaluate the performance new product.

1.1 Framework of Our Approach

Figure. 1 shows the framework of our approach. To sum up, the major contribution of our work as follows:

We infer the user goals by clustering, feedback sessions are proposed. Clustering the feedbacks can effectively reflect the user needs. Products can be improved effectively by providing ratings. So the user expectations can be obtained conveniently from the ratings. This can be very useful in new product development.

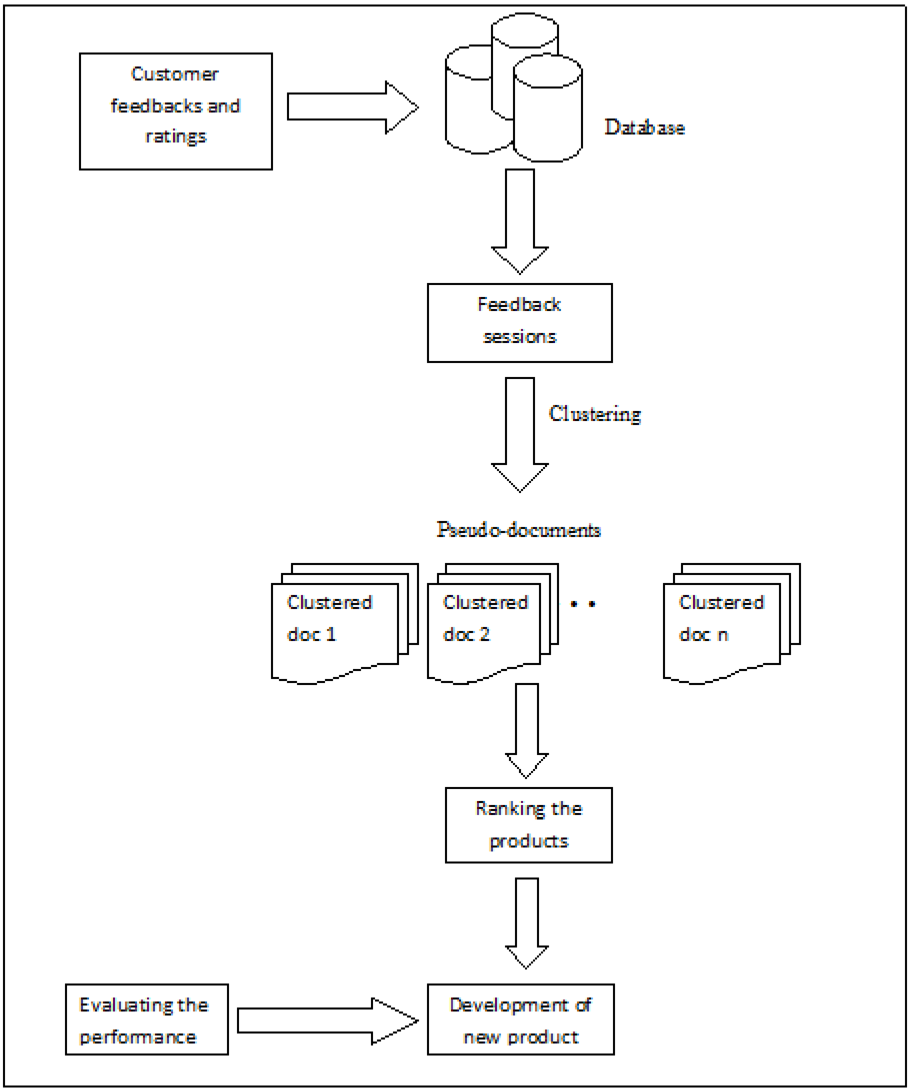


Figure.1 The framework of our approach

We propose the new product development. Through product development, it aims to meet the changing needs of its customers. Therefore organisation profits can be increase effectively.

We propose a method to analyse the customer behaviour. This helps in how the customer decision strategies differ between the products and how the marketing strategies more effectively reach the customer.

We propose the evaluation criterion to evaluate the performance of newly developed product. From this we can determine the user goals.

The rest of the paper is organised as follows: Related work is presented in section 2. The proposed feedback sessions and ratings are described in section 3. Section 4 describes the new product development. Analysing customer behaviour is presented in section 5. Section 6 deals with evaluation of product. Section 7 concludes the paper.

2. RELATED WORK

Many works about user search goals analysis have been investigated. In the case of search engine, the number of diverse user search goals for a query and depicting each goal with some keywords automatically [1]. Here the feedback sessions are formed with the series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs. [4], [8], [9], [12], [15] demonstrated the use of logs. In search engine the user goal can also be inferred by using the clickthrough data [7], [10]. It is more efficient to analyze the feedback sessions than to analyze the search results or clicked URLs directly. [13], [14] illustrated the work of inferring goals in search engine using clickthrough data. Users provide their own outcome and then answer questions that may be predictive of that outcome [2]. Models are constructed against the growing dataset that predict each user's behavioural outcome. Users pose their own questions that, when it answered by other users, then in the modeling process it become new independent variables.

Based on this we proposed feedback sessions by collecting the feedbacks about the products from the customers. Here we extend the work by forming the pseudo-documents. Pseudo-documents consist of clustered similar feedbacks. By clustering the feedbacks we can easily understand the user goals. So we can use these feedbacks in order to improve the products. Predicting the customer behaviour helps in the development of new product.

3. FEEDBACKS AND RATING

In this section, we first describe the feedback sessions and followed by ratings of the products. In this paper, we focus on inferring user search goals for a particular query. Each feedback session can tell what a user requires and what he/she does not care about. There are plenty of diverse feedbacks from the users therefore for inferring user search goals; it is more efficient to analyze the feedbacks.

Feedbacks are collected from every customer for each part of the product [5]. This helps in improving the product. Apart from feedbacks, ratings can also be given by the customers. Ratings are considered very important in comparing the product. Ratings benefit the customers and help them make informed purchasing decisions. Ratings increase confidence of consumer, and it enhance product visibility and ratings can dramatically increase sales. Ratings have the power to reach a large audience and be more influential than conventional marketing methods. One of the wonderful benefits of rating is that how the customers feel about brand, what they like and dislike about products and how can improve the overall product, which help products rank higher. H.-J Zeng [3] and J.-R Wen [6] illustrated the importance of clustering. Feedbacks for each part of product can be collected and can be effectively clustered by using k-means clustering algorithm which is effective and simple. We do not know the exact number of user search goals for each query, we set k to be different values and based on these values clustering is performed. User goals can also be predicted automatically. [11] and [17] shows the automatic identification of user goals. Compared to automatic identification collecting the customer feedbacks satisfy the user needs and expectations.

4. NEW PRODUCT DEVELOPMENT

Too many organisations suffer from customer amnesia, as though they have forgotten how to have routine conversations with their customers. When it comes to new product development, these organisations jump right to design of product, by assuming they know what customer expects, and then ships the finished product as soon as possible.

For successful reachability of product, find out what problems that organisation can solve for the customer before designing the product. Get early feedback on new product concepts from customers by showing them initial prototypes. These feedbacks can be collected from the customers by setting up the questions. Once the organisation has a system for collecting new product ideas and suggestions, it is easy to make up the product. Customers are a great resource for the product development feedback. Through product development, it aims to meet the changing needs of its customers and increases the customer total spend. Through offering more products and services it hopes to increase profits.

Customer feedbacks provide organization with valuable information that can be used to better position services or products in the marketplace. Still some companies are not asking the customers who buy their products and services what they want and need, while many companies do not incorporate customer suggestions into the product development process. Several reasons could explain that why some organizations do not include customer feedback into their product development and service improvement programs. Perhaps they do not realize the customer excellence is impossible to achieve without knowing or understanding what customers expects. Have they forgotten that the goal of collecting customer feedback regularly and proactively is to consistently exceed customer expectations? May be they are not aware that customer feedback programs can be used to create products or services that will ensure business success. Hence the customer suggestions are considered as very important in the development of new product.

5. ANALYSING USER BEHAVIOUR

Market research is often needed to ensure that what customer really wants. Analyzing customer behaviour helps organizations improve their marketing strategies by understanding how customers think and select between different alternatives. Customer motivation and decision strategies differ between products that differ in their level of importance. When the consumer behaviour and marketing strategy are intervened, marketers can expect success in their profit and sales, competitive sustainability and higher profit in the market place. The benefits of using consumer behaviour to create a marketing strategy are the knowledge marketer's gain about the needs and values of their target market. Once marketers understand this, their message will be delivered to the correct target in marketplace, resulting in an end sale. [18] introduced the machine science model for analyzing the customer behaviour.

Here we proposed the customer behaviour by analyzing the questions posed by the customers about the products. Customers can pose their questions. These questions are analysed by the investigator to check whether the question is suitable, if it is suitable then the question is selected and added by the investigator. For these questions other customers can also propose their answers or responses. [16] and [19] demonstrated the behavioural outcome of customers. These responses can be analyzed to predict the customer behaviour (Figure. 2). This effectively reflects the user needs and expectations which help in the new product development and improve the market sales.

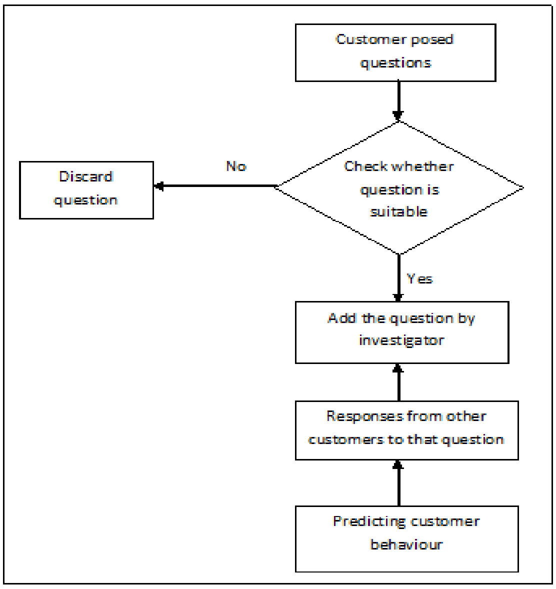

Customer posed questions
Check whether question is suitable
No
Discard question
Yes
Add the question by investigator
Responses from other customers to that question
Predicting customer behaviour

Figure. 2 predicting customer behaviour

6. EVALUATION OF PRODUCT

When developing a new product, an organisation should identify all the features. Determining the overall ranking of features by importance and relate the importance of each feature to its uniqueness reflects the importance of evaluation. The purpose of evaluation is to determine whether the outcome criteria have been met. It is done for the purpose of improvement. Some weaknesses can be found during evaluation.

Hence, evaluating the new product performance helps in identifying how far the product reaches successfully.

7. CONCLUSION

In this paper, user goals are inferred by clustering the feedbacks given by the customer. First the feedback sessions are proposed. Then the similar feedbacks are clustered to produce the pseudo-documents. Ratings which are given by the customers are collected. These feedbacks and ratings are used in the development of new product. Hence the knowledge and feedbacks from the customers has become important information. Customer behaviour has predicted by analysing the questions posed by the customers. The posed questions and responses are useful in predicting the user needs and expectations. Evaluating the new product helps in identifying the successful of product in market.

Through this, organization profit can be increased effectively. This helps in how the customer decision strategies differ between the products and how the marketing strategies more effectively reach the customer.

8. REFERENCES

[1] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin and Zhaohui Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions", IEEE transactions on knowledge and data engineering, march 2013.

[2] Josh C. Bongard, Paul D. H. Hines, Dylan Conger, Peter Hurd, and Zhenyu Lu, "Crowd sourcing Predictors of Behavioural Outcomes" IEEE transactions on knowledge and data engineering, 2013

[3] H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, "Learning to Cluster Web Search Results," Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '04), pp. 210-217, 2004.

[4] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004.

[5] B. Poblete and B.-Y Ricardo, "Query-Sets: Using Implicit Feedback and Query Patterns to Organize Web Documents," Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 41-50, 2008.

[6] J.-R Wen, J.-Y Nie, and H.-J Zhang, "Clustering User Queries of a Search Engine," Proc. Tenth Int'l Conf. World Wide Web (WWW '01), pp. 162-168, 2001.

[7] T. Joachims, "Evaluating Retrieval Performance Using Clickthrough Data," Text Mining, J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.

[8] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '00), pp. 407-416, 2000.

[9] X. Wang and C.-X Zhai, "Learn from Web Search Logs to Organize Search Results," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.

[10] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-Aware Query Suggestion by Mining Click-Through," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08), pp. 875-883, 2008.

[11] H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.

[12] C.-K Huang, L.-F Chien, and Y.-J Oyang, “Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs,” J. Am. Soc. for Information Science and Technology, vol. 54, no. 7, pp. 638-649, 2003.

[13] T. Joachims, “Optimizing Search Engines Using Clickthrough Data,” Proc. Eighth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD ’02), pp. 133-142, 2002.

[14] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay, “Accurately Interpreting Clickthrough Data as Implicit Feedback,” Proc. 28th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’05), pp. 154-161, 2005.

[15] R. Jones and K.L. Klinkner, “Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs,” Proc. 17th ACM Conf. Information and Knowledge Management (CIKM ’08), pp. 699-708, 2008.

[16] L. Barness, J. Opitz, and E. Gilbert-Barness, “Obesity: genetic, molecular, and environmental aspects,” American Journal of Medical Genetics Part A, vol. 143, no. 24, pp. 3016–3034, 2007.

[17] U. Lee, Z. Liu, and J. Cho, “Automatic Identification of User Goals in Web Search,” Proc. 14th Int’l Conf. World Wide Web (WWW ’05), pp. 391-400, 2005.

[18] J. Evans and A. Rzhetsky, “Machine science,” Science, vol. 329, no.5990, p. 399, 2010.

[19] T. Parsons, C. Power, S. Logan, and C. Summerbell, “Childhood predictors of adult obesity: a systematic review.” International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity, vol. 23, p. S1, 1999.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 2, 130 - 135, 2014

Analysis of Cryptographic Algorithms for Network Security

Kritika Acharya
DIT University
Dehradun, India

Manisha Sajwan
DIT University
Dehradun, India

Sanjay Bhargava
DIT University
Dehradun, India

Abstract: Cryptography plays a major role in securing data. It is used to ensure that the contents of a message are confidentially transmitted and would not be altered. Network security is most vital component in information security as it refers to all hardware and software function, characteristics, features, operational procedures, accountability, access control, and administrative and management policy. Cryptography is central to IT security challenges, since it underpins privacy, confidentiality and identity, which together provide the fundamentals for trusted e-commerce and secure communication. There is a broad range of cryptographic algorithms that are used for securing networks and presently continuous researches on the new cryptographic algorithms are going on for evolving more advanced techniques for secures communication.

Keywords: Cryptography, plain text, cipher text, encryption, decryption, network security.

1. INTRODUCTION

The building blocks of computer security are cryptographically-based mechanism. Cryptography can be applied anywhere in the TCP/IP stack, though it is not common at physical layer. Cryptography is also used in complicated protocols that help to achieve different security services, thus called security protocols. The main feature of the encryption/decryption program implementation is the generation of the encryption key [1].

1.1 Basic Terms Used in Cryptography

1.1.1 Plain Text

The original message that the person wishes to communicate with the other is defined as Plain the original message that the person wishes to communicate with the other is defined as Plain Text. In cryptography the actual message that has to be send to the other end is given a special name as Plain Text. For example, Alice is a person wishes to send “Hello Friend how are you” message to the person Bob. Here “Hello Friend how are you” is a plain text message.

1.1.2 Cipher Text

The message that cannot be understood by anyone or meaningless message is what we call as Cipher Text. In Cryptography the original message is transformed into non readable message before the transmission of actual message. For example, “Ajd672#@91ukl8*^5%” is a Cipher Text produced for “Hello Friend how are you”.

1.1.3 Key

A specific string of data that is used to encrypt and decrypt messages, documents or other types of electronic data.. Keys have varying levels of strength. Keys having higher numbers of bits are theoretically tougher to break because there are more possible permutations of data bits. (Since bits are binary, the number of possible permutations for a key of x bits is 2x.) The specific way a key is used depends on whether it's used with asymmetric or symmetric cryptography.

1.1.4 Encryption

A process of converting Plain Text into Cipher Text is called as Encryption. Cryptography uses the encryption technique to send confidential messages through an insecure channel. The process of encryption requires two things- an encryption algorithm and a key. An encryption algorithm means the technique that has been used in encryption. Encryption takes place at the sender side.

1.1.5 Decryption

A reverse process of encryption is called as Decryption. It is a process of converting Cipher Text into Plain Text. Cryptography uses the decryption technique at the receiver side to obtain the original message from non-readable message (Cipher Text). The process of decryption requires two things- a Decryption algorithm and a key. A Decryption algorithm means the technique that has been used in Decryption. Generally the encryption and decryption algorithm are same.

Now a day, cryptography has many commercial applications. If we are protecting confidential information then cryptography is provide high level of privacy of individuals and groups. However, the main purpose of the cryptography is used not only to provide confidentiality, but also to provide solutions for other problems like: data integrity, authentication, non-repudiation. Cryptography is the methods that allow information to be sent in a secure from in such a way that the only receiver able to retrieve this information. Cryptography not only protects data from theft or alteration, but can also be used for user authentication. It is necessary to apply effective encryption/decryption methods to enhance data security. Cryptography provides a number of security goals to ensure the privacy of data, non-alteration of data etc[2].

1.2 Goals of Cryptography

1.2.1 Confidentiality

Ensures that no one can read the message except the intended receiver.

1.2.2 Authentication

Mechanism to realize authentic communication i.e. the process of proving one’s identity.

1.2.3 Integrity

Assuming the receiver that the received message has not been altered in any way from the original.

1.2.4 *Non-Repudiation*

Ensures that neither the sender nor the receiver of message should be able to deny the transmission.

1.2.5 *Access Control*

Only the authorized parties are able to access the given information.

Network security involves the authorization of access to data in a network, which is controlled by the network administrator. The initial encrypted data is referred to as plain text. It is encrypted into cipher text, which will in turn be decrypted into usable plain text. Cryptographic algorithms are categorized based on the number of key that are employed for encryption and decryption[4].

1.3 Three Cryptographic Schemes

1.3.1 *Secret Key Cryptography Or Symmetric Cryptography*

Uses a single key for both encryption and decryption.

1.3.2 *Public Key Cryptography Or Asymmetric Cryptography*

Uses one key for encryption and another for decryption[2].

1.3.3 *Hash Function*

Uses a mathematical transformation to irreversibly "encrypt" information[14].

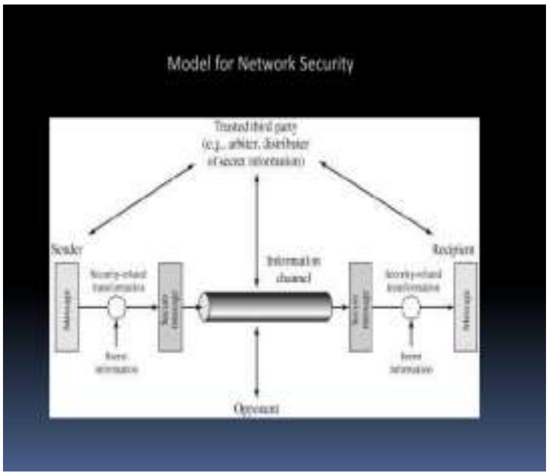


Figure 1. Model For network security

2. METHODOLOGY

Before implementing an encryption algorithm, we need to understand the principle behind the encryption i.e. to secure data held within a message or file and to ensure that the data is unreadable to others. The most important type of the encryption type is the symmetric key encryption. In the symmetric key encryption both for the encryption and decryption process the same key is used. Hence the secrecy of the key is maintained and it is kept private. It works with high speed. The symmetric key encryption takes place in two methodologies either as the block ciphers or as the stream ciphers. One of the main advantages of using the symmetric key encryption is that the computational power to this encryption technique is small. The keys for this are unique or there exists a simple transformation between the two keys[17].

Asymmetric key encryption is the technique in which the keys are different for the encryption and the decryption process. They are also known as the public key encryption. Public key methods are important because they can be used for transmitting encryption keys or other data securely even when the both the users have no opportunity to agree on a secret key in private, Algorithm[18]. Asymmetric algorithms are generally slow and it is impractical to use them to encrypt large amounts of data. The keys used in public-key encryption algorithms are usually much longer that improves the security of the data being transmitted[4].

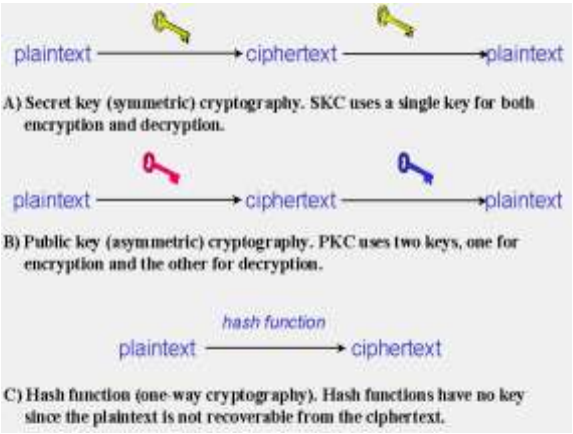

plaintext
ciphertext
plaintext
A) Secret key (symmetric) cryptography. SKC uses a single key for both encryption and decryption.
plaintext
ciphertext
plaintext
B) Public key (asymmetric) cryptography. PKC uses two keys, one for encryption and the other for decryption.
hash function
plaintext
ciphertext
C) Hash function (one-way cryptography). Hash functions have no key since the plaintext is not recoverable from the ciphertext.

Figure 2. Cryptographic Schemes

2.1 How Encryption Works in Cryptography

Encryption is not just a tool for spies and hackers, it can be a valuable asset even in the business world. For example, say you're an engineer for a company like Beyond the Office Door, an office furniture company that designs adjustable desks, and you just came up with a fantastic new adjustable desk design that will blow the world away. You can be pretty sure that your email is secure when sending information, but is "pretty sure" good enough when you're sending information on a new prototype adjustable desk? It's not, and thus it would be a perfect time for encryption to be used in the business world. And of course, there are many other valuable applications for encryption that are more mundane than trade secrets, like financial data, medical or legal information and so on.

The easy part of encryption is applying a mathematical function to the plaintext and converting it to an encrypted cipher. The harder part is to ensure that the people who are supposed to decipher this message can do so with ease, yet only those authorized are able to decipher it. We of course also have to establish the legitimacy of the mathematical function used to make sure that it is sufficiently complex and mathematically sound to give us a high degree of safety[5].

2.2 Classification Of Encryption Schemes

2.2.1 *Symmetric Key Encryption*

2.2.1.1 *DES(Data Encryption Standard)*

DES is a symmetric block cipher developed by IBM. The algorithm uses a 56-bit key to encipher/decipher a 64-bit block of data. The key is always presented as a 64-bit block, every 8th bit of which is ignored. However, it is usual to set each 8th bit so that each group of 8 bits has an odd number of bits set to 1.

2.2.1.2 *Triple DES(3DES)*

3DES is an enhancement of DES; it is 64 bit block size with 192 bits key size. In this standard the encryption method is similar to the one in the original DES but applied 3 times to increase the encryption level and the average safe time. It is a known fact that 3DES is slower than other block cipher methods[3].

2.2.1.3 AES

AES is a block cipher .It has variable key length of 128, 192, or 256 bits; default 256. It encrypts data blocks of 128 bits in 10, 12 and 14 round depending on the key size[16]. AES encryption is fast and flexible; it can be implemented on various platforms especially in small devices. Also, AES has been carefully tested for many security applications[11].

2.2.1.4 BlowFish

Blowfish algorithm is the important type of the symmetric key encryption that has a 64 bit block size and a variable key length from 32 bits to 448 bits in general. Since the key size is larger it is complex to break the code in the blowfish algorithm. Moreover it is vulnerable to all the attacks except the weak key class attack.

2.2.1.5 RC4

RC4 is recognized as the most commonly utilized stream cipher in the world of cryptography. RC4 has a use in both encryption and decryption while the data stream undergoes XOR together with a series of generated keys. It takes in keys of random lengths and this is known as a producer of pseudo arbitrary numbers.The output is then XORed together with the stream of data in order to generate a newly-encrypted data.

2.2.2 Asymmetric Key Encryption

2.2.2.1 RSA

Rivest-Shamir-Adleman is the most commonly used public key encryption algorithm. It can be used to send an encrypted message without a separate exchange of secret keys. It can also be used to sign a message. In RSA, this asymmetry is based on the practical difficulty of factoring the product of two large prime numbers, the factoring problem. RSA computation occurs with integers modulo n = p * q, for two large secret primes p, q. To encrypt a message m, it is exponentiated with a small public exponent e. For decryption, the recipient of the cipher text c = me (mod n) computes the multiplicative reverse d = e-1 (mod (p-1)*(q-1)) (we require that e is selected suitably for it to exist) and obtains cd = m e * d = m (mod n). The key size should be greater than 1024 bits for a reasonable level of security.

2.2.2.2 Diffie-Hellman Algorithm

The Diffie–Hellman key exchange method allows two parties that have no prior knowledge of each other to jointly establish a shared secret key over an insecure communications channel. This key can then be used to encrypt subsequent communications using a symmetric key cipher. The Diffie-Hellman protocol is generally considered to be secure when an appropriate mathematical group is used[10].

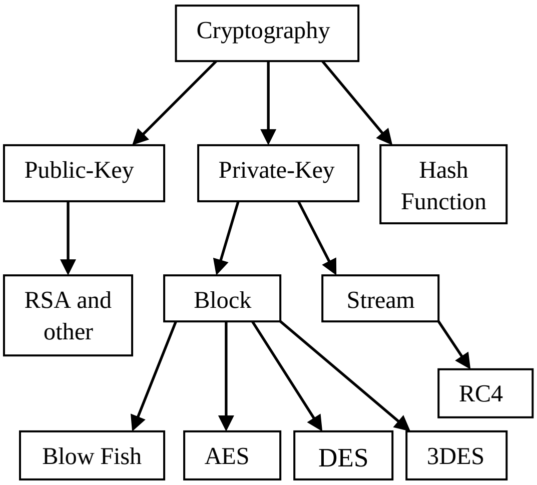


Figure 3. Model for Cryptography

2.3 Quantum Cryptography: A New Approach To Security

Quantum cryptography is a technology in which two parties can secure network communications by applying the phenomena of quantum physics. The security of these transmissions is based on the inviolability of the laws of quantum mechanics. Quantum cryptography is only used to produce and distribute a key, not to transmit any message data. Quantum cryptography is different from traditional cryptographic systems in that it relies more on physics, rather than mathematics, as a key aspect of its security model. Quantum cryptography uses our current knowledge of physics to develop a cryptosystem that is not able to be defeated - that is, one that is completely secure against being compromised without knowledge of the sender or the receiver of the messages. The genius of quantum cryptography is that it solves the problem of key distribution. A user can suggest a key by sending a series of photons with random polarizations. This Sequence can then be used to generate a sequence of numbers. The process is known as quantum key distribution. If the key is intercepted by an eavesdropper, this can be detected and it is of no consequence, since it is only a set of random bits and can be discarded. The sender can then transmit another key. Once a key has been securely received, it can be used to encrypt a message that can be transmitted by conventional means: telephone, e-mail, or regular postal mail[6].

2.4 Selection Of Right Cryptographic Scheme

The selection of right cryptographic technique relies on following constraints:

2.4.1 Time

How much time will be needed for encrypting and decrypting the data and how much time is need to fulfill the pre-requisites before starting an encryption how much time is need to fulfill the pre-requisites before starting an encryption.

2.4.2 Memory

How much memory will be need especially in case of small devices like PDAs, smart cards, RFID tags.

2.4.3 Security

Selected encryption scheme should meet the confidentiality, integrity (authentication, non-repudiation) and availability.

2.4.4 Nature Of Data

Nature of data means the communicating information is how much confidential or important. If the information is small in size and not too much important; then any encryption scheme is suitable. If information is highly secret or important then joint hybrid combination of symmetric + asymmetric scheme will be suitable[13].

2.4.5 Type Of Data

In case of video data the privacy is more valuable and considerable constraint. If the data is small and in video format the previous described constrains (Time, memory, security) suggest the use of asymmetric scheme but this selection is not sufficient because the third party especially in case of Identity based Public Key Cryptography (ID-PKC) can view the video clip as they have all information (key(s), encrypted data). So in this case the privacy is nothing. That's why the type of data constraint is highly important constraint which should not be neglected in case of right selection of cryptographic scheme. If data type is confidential multimedia (personal video clip) then the symmetric scheme is good but hybrid encryption method (symmetric + asymmetric) can provide all security objectives[12].

2.5 Performance Factors

Various important factors on which performance of cryptographic algorithms depend are:

2.5.1 Tunability

It could be very desirable to be able to dynamically define the encrypted part and the encryption parameters with respect to different applications and requirements. Static definition of encrypted part and encrypted parameters limits the usability of the scheme to a restricted set of applications.

2.5.2 Computational Speed

In many real-time applications, it is important that the encryption and decryption algorithms are fast enough to meet real time requirements.

2.5.3 Key Length Value

In the encryption methodologies the key management is the important aspect that shows how the data is encrypted. The image loss the encryption ratio is based on this key length. The symmetric algorithm uses a variable key length which is of the longer. Hence, the key management is a considerable aspect in encryption processing.

2.5.4 Encryption Ratio

The encryption ratio is the measure of the amount of data that is to be encrypted. Encryption ratio should be minimized to reduce the complexity on computation[8].

2.5.5 Security Issues

Cryptographic security defines whether encryption scheme is secure against brute force and different plaintext-cipher text attack? For highly valuable multimedia application, it is really important that the encryption scheme should satisfy cryptographic security. In our analysis we measure cryptographic security in three levels: low, medium and high[9].

Table 1. Comparison table for various cryptographic algorithms

Algorithm	**Key Size(s)**	**Speed**	**Speed Depends On Key?**	**Security**
DES	56 bits	Slow	Yes	Insecure
3DES	112/168 bits	Very Slow	No	Moderately secure
AES	128, 192, 256 bits	Fast	Yes	Secure
BLOW-FISH	32-448 bits	Fast	No	Believed secured, but less attempted crypt-analysis than other algorithms
RC4	256 bytes	Very Fast	No	Moderately secure
RSA	1024 bits and above	Fast	Yes	Secure

2.6 Trend In Cryptographic Protocol

In this section we describe what we see as some of the emerging trends in cryptographic protocols. These trends present new challenges to protocol analysis

2.6.1 Greater Adaptability and Complexity:

Probably one of the most obvious trends is the increasing different kinds of environments that protocols must interoperate with. As networks handle more and more tasks in a potentially hostile environment, cryptographic protocols take on more and more responsibilities. As networking becomes more widespread, and different platforms must interoperate, we see protocols such as the Internet Key Exchange (IKE) protocol that not only must agree upon encryption keys, but on the algorithms that are to use the keys. Or, we may see protocols such as SET that must be able to process different types of credit card transactions.One way of attempting to meet this challenge is to increase the complexity of the protocol. This of course, not only makes verification but implementation more difficult as well, and as a result there is always resistance to this approach. However, the tendency to greater complexity will always be there, and it will ultimately have to be met at least part of the way by anyone who is attempting to perform any type of security analysis[7].

2.6.2 Adoption of New Types of Cryptographic Primitives

In general, it is it is accepted that a conservative approach to algorithm is best when designing cryptographic protocols;

only tried and true algorithms should be used. But, as the field matures the number of algorithms that are considered to have received enough scrutiny has increased. Moreover, as computing power increases, algorithm that were once considered prohibitively expensive have become easier to implement, while others, such as DES, are widely regarded as no longer providing adequate security[20].

2.6.3 New Types of Threats:

In the early years of computer security, much of the threat analysis was hypothetical, and focused on attacks in which there would be a clear(usually monetary) gain for the attacker. Now, with more experience, we see that there are other types of attacks, most of them related to denial of service, that can prevent a network from fulfilling its functions. Many denial of service attacks can be countered by good resource management. But sound protocol design can do much to help, for example by keeping a responder from committing its resources to communicating with an initiator until it has adequate assurance that it knows who it's talking to. This can be a delicate problem however, since many of the techniques used for authentication themselves require commitment of resources, and since the decision of how much resources to commit, and when, can be very implementation-dependent. Successful analysis will depend to some extent on the ability to compare the resources expended by an attacker to the resources expended by a defender.

Other threats, such as traffic analysis, focus on problems that are not really an issue until adequate cryptographic protection for communication secrecy has already been attained. Protection against traffic analysis is one of these. Even when encryption is used source and destination of message traffic is not hidden, and it can be possible for an observer to learn much from this alone. A number of different systems have been developed that attempt to solve this problem with varying degrees of completeness. However, without some ability to evaluate and compare the degree of protection offered by these systems, it is difficult to assess what amount and kind of security they offer. Such analysis methods should take statistical techniques into account, since much traffic analysis depends on statistical analysis[19].

A somewhat different type of threat emerges when we look at electronic commerce protocols. In this type of protocol, the parties involved participate in a transaction that results in certain levels of payoff to each principal involved. Moreover, the protocol may either depend upon or try to guarantee liveness or fairness properties as well as safety properties. A principal may try to cheat by trying to increase its payoff at the expense of those of other parties, but will not engage in behavior that will result in a lowering of its payoff, or put it at a disadvantage with respect to the others[15].

3. CONCLUSION

Cryptography is an emerging technology which is important for network security. Some well-known cryptographic algorithms have been analyzed in this paper to demonstrate the basic differences between the existing encryption techniques. Regardless of the mathematical theory behind an algorithm, the best algorithm are those that are well-known and well-documented because they are well-tested and well-studied. In-fact time is the only true test of good cryptography, any cryptographic scheme that stays in use year after year is most likely good one. The strength of cryptography lies in the choice of the key; longer key resist attack better than shorter keys. No one can guarantee 100% security. But we can work toward 100% risk acceptance. Fraud exists in current commerce systems: cash can be counterfeited, checks altered, credit card numbers stolen. A good cryptographic system strikes a balance between what is possible and what is acceptable. Thus considerable research effort is still required for secured communication.

4. REFERENCES

[1] William Stallings "Network Security Essentials (Applications and Standards)", Pearson Education, 2004.

[2] W. Stallings. "Cryptography and Network Security", Prentice Hall, 1995.

[3] National Bureau of Standards, "Data Encryption Standard," FIPS Publication 46, 1977.

[4] E. Thambiraja, G. Ramesh, Dr. R. Umarani, "A Survey on Various Most Common Encryption Techniques" International Journal of Advanced Research in Computer Science and Software Engineering, VOL. 2, Issue 7 July 2012, Page 226-233.

[5] Sumedha Kaushik, Ankur Singhal, "Network Security Using Cryptographic Techniques" International Journal of Advanced Research in Computer Science and Software Engineering, VOL.2, Issue 12 December 2012, Page 105-107.

[6] Vishwa gupta, Gajendra Singh, Ravindra Gupta, "Advance cryptography algorithm for improving data security" International Journal of Advanced Research in Computer Science and Software Engineering, VOL.2, Issue 1 January 2012.

[7] Nagamalleswara Rao. Dasari, Vuda Sreenivasarao, "PERFORMANCE OF MULTI SERVER AUTHENTICATION AND KEY AGREEMENT WITH USER PROTECTION IN NETWORK SECURITY" International Journal on Computer Science and Engineering, VOL.2, Issue 05 2010, Page 1705-1712.

[8] AL. Jeeva, Dr. V. Palanisamy, K. Kanagaram, "COMPARATIVE ANALYSIS OF PERFORMANCE EFFICIENCY AND SECURITY MEASURES OF SOME ENCRYPTION ALGORITHMS" International Journal of Engineering Research and Applications (IJERA), VOL.2, Issue 3,May-Jun 2012, Page 3033-3037.

[9] G. Ramesh, R. Umarani, "Performance Analysis of Most Common Encryption Algorithms on Different Web Browsers "I.J. Information Technology and Computer Science, Issue Nov 2012, Page 60-66.

[10] Zirra Peter Buba & Gregory Maksha Wajiga "Cryptographic Algorithms for Secure Data Communication "in International Journal of Computer Science and Security IJCSS, Volume no 5, Issue 2.

[11] Daemen, J., and Rijmen, V. "Rijndael: The Advanced Encryption Standard." Dr. Dobb's Journal, March 2001.

[12] Pranay Meshram,Pratibha Bhaisare, S.J.Karale,",comparative study of selective encryption algorithm for wireless adhoc network" ,IJREAS Volume 2, Issue 2 , in International Journal of Research in Engineering & Applied Sciences.

[13] Yudhvir Singh, Yogesh Chaba, ―Information Theory test based Performance Evaluation of Cryptographic Techniques , International Journal of Information

Technology and Knowledge Management, Vol 1,No.2,2008 , pp. 475-483.

[14] A. Menezes, P. van Oorschot, S. Vanstone, Algorithm 9.53 Secure Hash Algorithm - revised (SHA-1), Handbook of Applied Cryptography, CRC Press, 1997.

[15] M. Merkow, J. Breithaupt, J. Breithaupt, The Complete Guide to Internet Security, AMACOM, 2000.

[16] Punita Mellu & Sitender Mali, "AES: Asymmetric key cryptographic System" International Journal of Information Technology and Knowledge Management, 2011, Vol, No. 4 pp. 113-117.

[17] Suhaila Orner Sharif, S.P. Mansoor, ―"Performance analysis of Stream and Block cipher algorithms", 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 2010.

[18] Murat Fiskiran, Ruby B. Lee, "Workload Characterization of Elliptic Curve Cryptography and other Network Security Algorithms for Constrained Environments" IEEE International Workshop on Workload Characterization, 2002. WWC-5. 2002.

[19] Othman O. Khalifa, MD Rafiqul Islam, S. Khan and Mohammed S. Shebani,"Communication Cryptography", 2004 RF and Microwave Conference, Oct 5-6, Subang, Selangor, Malaysia.

[20] Mohamed A.Haleem, Chetan N. Mathur, R. Chandramouli, K. P. Subbalakshmi, "Opportunistic Encryption: A tradeoff between Security and Throughput in Wireless Network" IEEE Transactions on Dependable and secure computing, vol. 4, no. 3.