International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 01 - 06, 2015, ISSN:- 2319–8656

Arabic Numerals Recognition Using Linear Correlation Algorithm in Two Dimension

Gafar Zen Alabdeen Salh
Department of IT
Faculty of Computers and IT
King Abdulaziz University
Khulais, Saudi Arabia

Abdelmajid Hassan Mansour
Department of IT
Faculty of Computers and IT
King Abdulaziz University
Khulais, Saudi Arabia

Abdelhafeez hamid mohammed
Department of CS
Faculty of CS and IT
Kassala University
Kassala, Sudan

Abstract: The process of handwriting recognition and text is an important field which have a large role in many applications, including the identification of manually written digits on checks and documents, also the recognition of the postal addresses using the technology of Optical Character Recognition (OCR), and etc. The aim of this paper is to using the linear correlation algorithms in two dimensions for the purpose of Arabic numerals (Indian) recognition (0-1-2-3-4-5-6-7-8-9). So as to overcome the problems of documents that stored in the form of image. And searching or editing it, In order to recognize the Arabic digits on them.

Keywords: Optical Character Recognition; Handwriting; Image Processing; Pattern Recognition; artificial Neural Networks

1. INTRODUCTION

The handwriting recognition refers to the identification of written characters. Handwriting recognition has been become a very important and useful research area in recent years for the ease of access of many applications [2].

Numeral recognition refers to the process of translating images of handwritten, typewritten, or printed digits into a format understood by the user for the purpose of editing, indexing, searching and reduction in storage size .Number recognition can be online or offline. In online number recognition, data are captured during the writing process with the help of special pen and an electronic interface. Offline documents are scanned images of prewritten text, generally on sheet or paper [2].

Handwritten character recognition is a field of image processing as well as pattern recognition. There are two approaches for the pattern recognition such as statistical and structural. In statistical approach, the characteristic measurements of the input data is generated on the statistical basis and is assigned to one of the n classes. The structural description of the object is based on the interconnections and interrelationships of features of input data. In general, both approaches are widely used in the pattern recognition. Since the handwriting of different writers is different, building a general recognition system that would recognize all characters with good reliability is not possible in every application. Thus recognition systems are developed to achieve reliable performances to the specific applications [1]. The problem is quite complex, and even now there is no single approach that solves it both efficiently and completely in all settings [4].

In the handwriting recognition process, an image containing text must be appropriately supplied and pre-processed. Next, the text must either undergo segmentation or feature extraction. Small processed pieces of the text will be the result, and these must undergo recognition by the system. Finally, contextual information should be applied to the recognized symbols to verify the result [4].

The advance of handwriting processing results from a combination of various elements, for example: improvements in the recognition rates, the use of complex systems to integrate various kinds of information, and new technologies such as high quality high speed scanners and cheaper and more powerful CPUs. Some handwriting recognition system allows us to input our handwriting into the system. This can be done either by controlling a mouse or using a third-party drawing tablet [5].

2. RELATED WORKS

Handwritten recognition is becoming more and more important in the modern world. It helps humans ease their jobs and solve more complex problems. There many studies about the Handwritten recognition. Sumedha B. Hallale, Geeta D. Salunke was designed a back propagated neural network and trained it with a set of handwritten digits. The average success rates of recognition of all digits are 91.2% [1]. a system that recognizes an English numeral, given by the user, which is already trained on the features of the numbers to be recognized using NNT (Neural network toolbox) was proposed by Amritpal kaur, Madhavi Arora [2].

Satish Lagudu, CH.V.Sarma was proposed creation of a new handwritten language recognition method. They deals with recognition of isolated handwritten characters and words using Hybrid Particle swarm Optimization and Back Propagation Algorithm [4]. MALOTHU NAGU, N VIJAY SHANKAR, K.ANNAPURNA was proposed Two techniques, are Pattern Recognition and Artificial Neural Network (ANN). Both techniques are defined and different methods for each technique is also discussed [5]. Poornima G Patil, Ravindra S Hegadi was introduced the handwritten signatures images from a standard database are preprocessed and are decomposed using wavelets. The wavelet approximation and detail coefficients in three directions are subjected to principal component analysis and are used to train the SVM classifier using a linear kernel and a nonlinear kernel which is Gaussian Radial Basis Function kernel [6].

Sameer Singh, Adnan Amin was introduced the automatic recognition of hand-printed Latin characters using artificial neural networks in combination with conventional techniques [7]. A mixture model by concurrently performing global data partition and local linear PCA. The partition is optimal or near

optimal, which is realized by a soft competition algorithm called &neural gas' was proposed by Bailing Zhang, Minyue Fu, and Hong Yan [8].

Deepika Wadhwa, Karun Verma was presented an online handwritten Hindi numeral recognition system using Support Vector Machines (SVM) [9]. Ashish Gupta, Bhagwat Kakde was dealing with the unique method to identify cursive handwriting detection using artificial neural network (ANN) [10]. A method for Offline Handwritten Arabic Numerals Recognition with the use of Classifier and Feature Extraction Techniques was proposed Gita Sinha, Jitendra kumar [11]. Sabri A. Mahmoud, Sunday O. Olatunji was proposed a technique for handwritten Arabic (Indian) numerals recognition using multi-span features is presented. Angle, ring, horizontal, and vertical span features are used [12].

3. PROPOSED SCHEME

The proposed work uses the method of linear correlation algorithms in two dimensions for the purpose of recognizing Arabic numerals (Indian). In order to solve the problems of documents that stored in the form of image, and searching or editing it, and recognizing the Arabic digits on it. This scheme depends on the Method of automatic recognition of characters and Optical Character Recognition (OCR), they divided the process of automated reading into five phases, as shown in Figure 1.

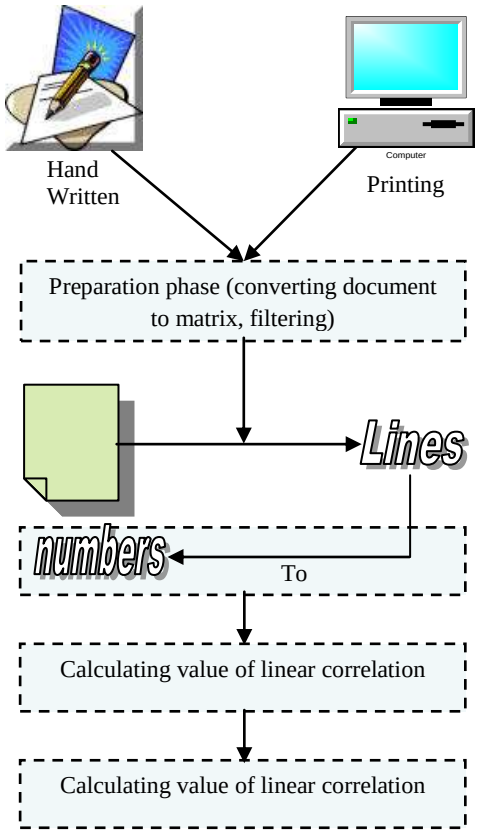

Figure. 1 Description of the system phase

The general structure of the proposed scheme is described in a diagram, representing the fifth steps of automated reading as shown in Figure 2.

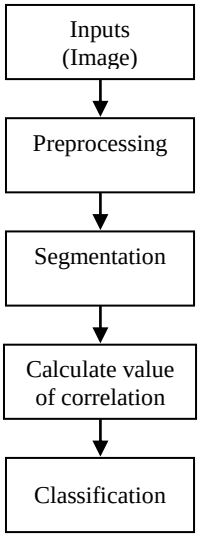

Figure. 2 General structure of the system

3.1 inputs stage

Is the first step in the algorithm, the system takes the original image that needed to read, from the scanner or from computer storage, this process can be represented in follow chart, as shown in Figure 3.

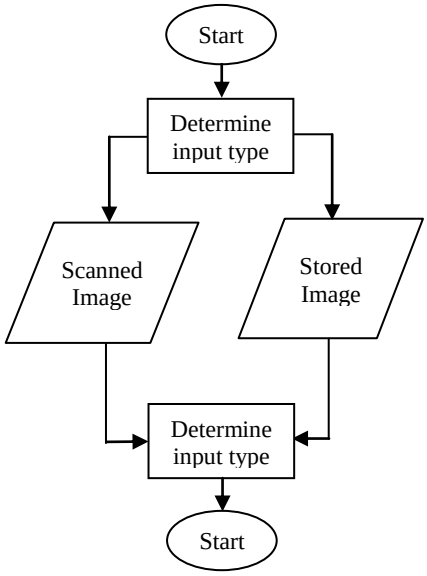

Figure 3. Input stage flow chart

3.2 Preprocessing stage:

Is the second step in the algorithm, the system get the original image that stored in a computer storage, then starting the preprocessing according to the following steps:

1) Checking the image, is colored or not for transforming it to gradient.
2) Converting the image to thresholding binary matrix in two dimensions, the symbol (0) representing the white squares, while the symbol (1) representing the black squares.
3) Determine the size and refine the distortions of the image and their impurities which may be associated with it.

4) This process can be represented in follow chart, as shown in Figure 4.

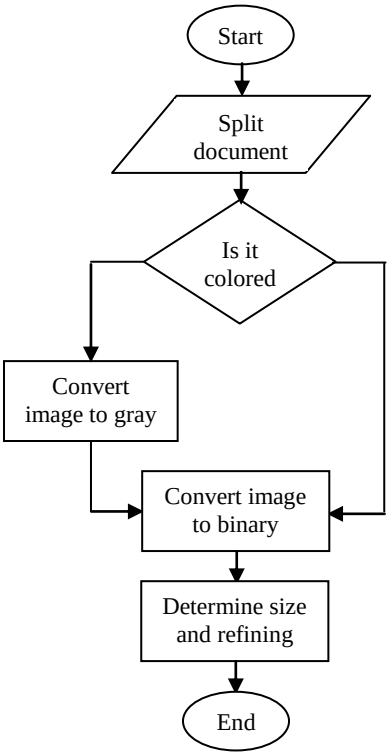

Figure 4. Preprocessing stage flow chart

3.3 Segmentation stage:

Is the third step in the algorithm, the inputs of the system is two dimensional matrix, as shown in Figure 5.

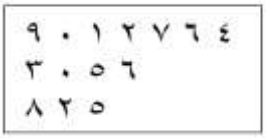

Figure 5. Inputs of the system

The segmentation stage moves through the following steps:

1) The document divided into lines by using histogram, then calculating the number of dots in each horizontal pointed row. Then in normal state we noticed that the number of dots is equal to a zero or close to zero in some horizontal rows. Here the system detects that means this row is between two lines, (above this row and below it). Then after repeating this process on both lines, the reader may distinguished lines of the document, and divide it, as shown in Figure 6.

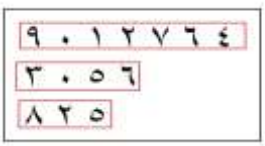

Figure 6. Dividing inputs into lines

2) Dividing the lines of the document into digits, according to the shape of the digits, based on rules and information that owned by the system, as shown in Figure 7.

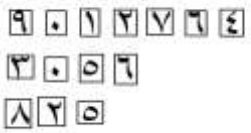

Figure 7. Dividing inputs lines into digits

3) Extracting the features by collecting the dots in each row separately, and also to the columns, then studying and analyzing the characteristics such as digit height and width, in preparation to identify the number.

4) The three steps can be represented in follow chart, as shown in Figure 8.

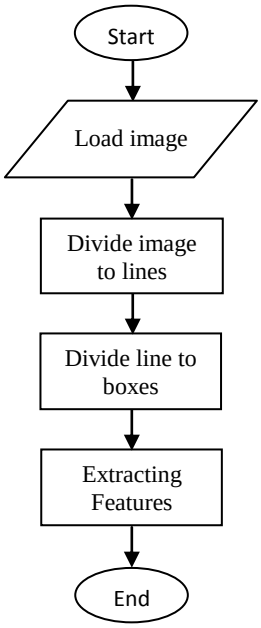

Figure 8. Segmentation stage flow chart

3.4 Calculating correlation value stage:

Is the fourth step in the algorithm, the system calculate correlation value of the digit matrix that want to recognize it and the template matrix by using the following equation.

$$r = \frac{\sum_m \sum_n (A_{mn} - \overline{A})(B_{mn} - \overline{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \overline{A})^2 (B_{mn} - \overline{B})^2 \right)}}$$

Where:

r = correlation value.

A = initial matrix (for digit want to identify it).

B = template matrix (for stored digit).

$\overline{A}$ = mean of the initial matrix (for digit want to identify it).

$\overline{B}$ = mean of the template matrix (for stored digit).

The steps of calculating the correlation value is represented in flow chart, as shown in Figure 9.

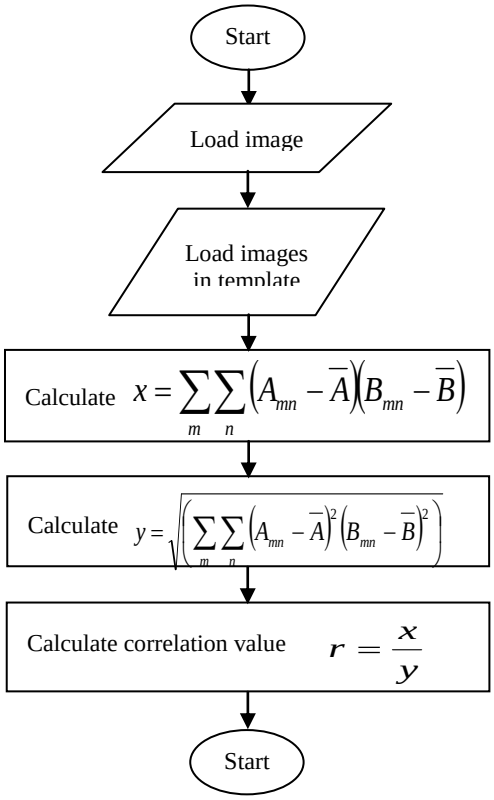

Figure 8. Calculating correlation stage flow chart

3.5 Classification stage:

Is the fifth step in the algorithm, the system compare the result of correlation value that calculated from the entered image with the stored message in the template, the system take the digit that is corresponding to the max value calculated to the correlation, and consider it is the digit from the entered image, then deliver the digits in text file.

4. EVALUATING THE RESULTS:

The system was tested on 10 images, each image contains all set of the Arabic numerals (10 digits). The result was analyzed by statistical analysis software (SPSS) to calculate the false rejection rate, which is considered a good measure to assess the Pattern Recognition Systems, as shown in Tables.

Table 1. Descriptions of the test statistics

number / Image	0	1	2	3	4	5	6	7	8	9
One	1	1	1	1	1	1	1	1	1	1
Two	1	1	1	0	1	1	1	1	0	0
Three	1	1	1	0	1	1	0	1	1	1
Four	1	1	1	1	1	1	1	1	1	1
Five	1	1	1	1	1	1	1	1	0	0
Six	1	1	1	1	1	1	1	1	0	0
Seven	1	1	1	1	1	1	0	1	1	1
Eight	1	1	1	1	1	1	1	1	1	0
Nine	1	1	1	1	1	1	1	1	0	0
Ten	1	1	1	1	1	1	0	1	1	1

Table 2. Statistics for the digit (Zero)

		Frequency	Percent	Valid Percent	cumulative Percent
Valid	true Classification	10	100.0	100.0	100.0

Table 3. Statistics for the digit (One)

		Frequency	Percent	Valid Percent	cumulative Percent
Valid	true Classification	10	100.0	100.0	100.0

Table 4. Statistics for the digit (Two)

		Frequency	Percent	Valid Percent	cumulative Percent
Valid	true Classification	10	100.0	100.0	100.0

Table 5. Statistics for the digit (Three)

		Frequency	Percent	Valid Percent	cumulative Percent
Valid	false Classification	2	20.0	20.0	20.0
	true Classification	8	80.0	80.0	80.0
	Total	10	100.0	100.0	

Table 6. Statistics for the digit (Four)

		Frequency	Percent	Valid Percent	cumulative Percent
Valid	true Classification	10	100.0	100.0	100.0

Table 7. Statistics for the digit (Five)

		Frequency	Percent	Valid Percent	cumulative Percent
Valid	true Classification	10	100.0	100.0	100.0

Table 8. Statistics for the digit (Six)

		Frequency	Percent	Valid Percent	cumulative Percent
Valid	false Classification	3	30.0	30.0	30.0
	true Classification	7	70.0	70.0	70.0
	Total	10	100.0	100.0	

Table 9. Statistics for the digit (Seven)

		Frequency	Percent	Valid Percent	cumulative Percent
Valid	true Classification	10	100.0	100.0	100.0

Table 10. Statistics for the digit (Eight)

		Frequency	Percent	Valid Percent	cumulative Percent
Valid	false Classification	4	40.0	40.0	40.0
	true Classification	6	60.0	60.0	100.0
	Total	10	100.0	100.0	

Table 11. Statistics for the digit (Nine)

		Frequency	Percent	Valid Percent	cumulative Percent
Valid	false Classification	5	50.0	50.0	50.0
	true Classification	5	50.0	50.0	100.0
	Total	10	100.0	100.0	

5. CONCULUSION

Through the statistics analysis in the tables above, we found that the system is competence to identify amount of images that contain Arabic digits in size of 20x20 pixels, as in the following points:

1) Rate of true classification of the digits (0, 1, 2, 4, 5, and 7) is 100% and the false classification rate is 0%.

2) Rate of true classification of the digit (3) is 80% and the false classification rate is 20%.

3) Rate of true classification of the digit (6) is 70% and the false classification rate is 30%.

4) Rate of true classification of the digit (8) is 60% and the false classification rate is 40%.

5) Rate of true classification of the digits (9) is 50% and the false classification rate is 50%.

6. REFERENCES

[1] Sumedha B. Hallale1, Geeta D. Salunke ," OFFLINE HANDWRITTEN DIGIT RECOGNITION USING NEURAL NETWORK" International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering(ijareeie), ISSN : 2278 – 8875, Vol. 2, Issue 9, September 2013, pp. 4373-4377.

[2] Amritpal kaur, Madhavi Arora, "Neural network based Numerical digits Recognization using NNT in Matlab", International Journal of Computer Science & Engineering Survey (IJCSES), Vol.4, No.5, October 2013, pp. 19-29.

[3] S.KNERR, L.PERSONNAZ, G.DREYFUS, "Handwritten Digit Recognition by Neural Networks with Single-Layer Training", IEEE TRANSACTIONS ON NEURAL NETWORKS, vol. 3, 962(1992), pp. 1-18.

[4] Satish Lagudu, CH.V.Sarma, "HAND WRITING RECOGNITION USING HYBRID PARTICLE SWARM OPTIMIZATION & BACK PROPAGATION ALGORITHM" , International Journal of Application or Innovation in Engineering & Management (IJAIEM), ISSN: 2319 – 4847, January 2013, Volume 2, Issue 1, pp. 75-81.

[5] MALOTHU NAGU, N VIJAY SHANKAR, K.ANNAPURNA, "A novel method for Handwritten Digit Recognition with Neural Networks", International Journal of Computer Science and Information Technologies (IJCSIT), ISSN:0975-9646, Vol. 2 (4) , 2011, pp. 1685-1692.

[6] Poornima G Patil, Ravindra S Hegadi, "Offline Handwritten Signatures Classification Using Wavelets and Support Vector Machines", International Journal of Engineering Science and Innovative Technology (IJESIT), ISSN: 2319-5967, Volume 2, Issue 4, July 2013, pp. 573-579.

[7] Sameer Singh, Adnan Amin, "Neural Network Recognition of Hand Printed Characters" Neural Computing and Applications, vol. 8, no. 1, 1999, pp. 67-76.

[8] Bailing Zhang, Minyue Fu, Hong Yan, "A nonlinear neural network model of mixture of local principal component analysis: application to handwritten digits recognition", the journal of the pattern recognition society, 34 (2001), pp. 203-214.

[9] Deepika Wadhwa, Karun Verma, "Online Handwriting Recognition of Hindi Numerals using SvmDeepika", International Journal of Computer Applications (0975 – 888), Volume 48– No.11, June 2012, pp. 13-17.

[10] Ashish Gupta, Bhagwat Kakde, "A Novel Approach for Cursive Handwriting Detection Using Artificial Neural Network", International Journal of Advanced Research in Computer Science and Software Engineering(ijarcsse), ISSN: 2277 128X, Volume 3, Issue 11, November 2013, pp. 674-680.

[11] Gita Sinha, Jitendra kumar, "Arabic numeral Recognition Using SVM Classifier", International Journal of

Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-2, Issue-5), pp. 62-67.

[12] Sabri A. Mahmoud, Sunday O. Olatunji, "HANDWRITTEN ARABIC NUMERALS RECOGNITION USING MULTI-SPAN FEATURES & SUPPORT VECTOR MACHINES", 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010), 978-1-4244-7166-9/10/$26.00 ©2010 IEEE, pp. 618-621.

[13] Anthony, M. and Bartlett, P. ,1999, " Neural Network Learning: Theoretical Foundations". Cambridge University Press, Cambridge, UK. 1999.

[14] Bishop, C. M. ,1995, "Neural Networks for Pattern Recognition". Oxford University Press, Oxford, UK. 1995.

[15] Martin H. Luerssen, Character Recognition with Neural Networks Flinders, University of South Australia.

International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 07 - 10, 2015, ISSN:- 2319–8656

Semantically Enriched Knowledge Extraction With Data Mining

Anuj Tiwari
Department of Civil Engineering
Indian Institute of Technology-Roorkee
India

P. Srujana
Computer Science and Engineering
CMR Technical Campus-Hyderabad
India

K. Rajesh
Computer Science and Engineering
CMR Technical Campus-Hyderabad
India

Abstract — **while data mining has enjoyed great popularity and success in recent years, Semantic web is shaping up as a next big step in the evolution of World Wide Web. It is the way web is growing as a smarter cyberspace. In field of Information and communication technology huge amount of data is available that need to be turned into knowledge. On the one side Data Mining is a nontrivial extraction of implicit, previously unknown and potentially useful knowledge from data in databases and on the other side Semantic web developing new platform to represent extracted knowledge in both the machine and human understandable format. The aim of this paper is to explore the concept of data mining in the context of semantics. Paper uses a basic input dataset with an open source software WEKA and a commercial one SAS for knowledge discovery; further this knowledge is represented in human understandable format with NLP (Natural Language Processing library) and in machine understandable format (RDF) with an indigenous algorithm implemented with java.**

Keywords— Data Mining; Semantic Web; Ontology; Knowledge; RDF; WEKA; SAS.

1. INTRODUCTION

We all are surrounded with a lot of data. Every time we watch television, we do any type of search on the internet, we swipe our ATM card etc more and more data is generated. In order to explore, analyze and discover valid, implicit, novel, understandable, potentially useful patterns, associations or relationships in large quantities of data a number of analytical tools are required that allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Data mining is the collection of methods that analyze data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both.

Valid: The patterns hold in general.
Novel: We did not know the pattern beforehand.
Useful: We can devise actions from the patterns.
Understandable: We can interpret and comprehend patterns.

Data mining deals with what kind of patterns can be mined. In this world of information and communication technology data is continuously growing like anything. This flood of data and sophisticated tools of data mining together very productive for business purpose where companies are interested in various patterns like purchase, educational, traffic, habits etc.

Data mining tasks are generally divided into two major categories. The objective of predictive tasks is to predict the values of a particular attributes based on the values of other attributes while Descriptive tasks derive patterns (correlations, trends, clusters, trajectories and anomalies) to summarize the underlying relationships in data.

Evolution of internet in last couple of decades brought a remarkable growth in the development of new technologies and applications, contributing to a historic transformation in the way we work, communicate, socialize, learn, create and share information, and organize the flow of people, ideas, and things around the globe. Being an extension of the existing web technology 'Semantic Web' is well recognized now as an effective infrastructure to enhance visibility of knowledge on the Web for humans and computers alike [1]. 'Semantic Web' enables the description of contents and services in machine-readable form, and enables annotating, discovering, publishing, advertising and composing services to be automated. It was developed based on Ontology, which is considered as the backbone of the Semantic Web [2]. Jasper and Uschold identify three major uses of semantic web and ontologies [3]:

(i) To assist in communication between human and computers,
(ii) To achieve interoperability (communication) among software systems, and
(iii) To improve the design and the quality of software systems.

2. METHODOLOGY

Methodology adopted for RDF generation is as follows:

Step 1: Input Data is retrieved from the DBMS.
Step 2: Open source data mining tool WEKA is used for data preprocessing and Nominal data set preparation.

Step 3: Nominal data is passed to commercial data mining tool SAS, regression is applied on that data.
Step 4: Output dataset is saved in database.
Step 6: Open NLP is a tool which is used to parse the data which is retrieved from the database.
Step 7: Triples are extracted from the sentences using Stanford NLP parser tool.
Step 8: RDF is generated using APACHE JENA.

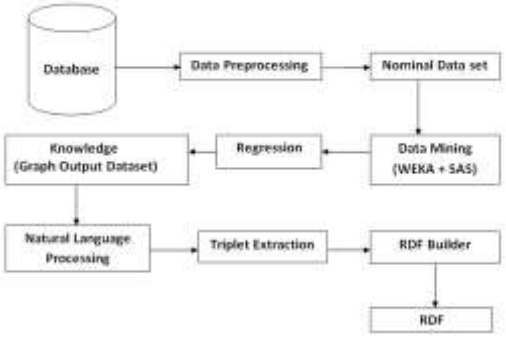

Fig. 1. Flow digram dipecting adopted methodology.

Standard database management system (like mySql) is used to save and mange a larger database. This larger database is preprocessed to generate nominal data set which in turn mined with open source data mining tool WEKA and commercial data mining tool SAS. WEKA generates graphical results and with SAS tabular results are obtained. Here regression is used as data mining method. Extracted knowledge is processed with Natural Language Processing (NLP) library and human interactive triplets are generated. RDF Builder step process these triplet and generate machine understandable RDF data set.

3. DESCRIPTION

A. Input Dataset

Data is stored in MS ACCESS database. It is retrieved as ".CSV" file. The data file is shown below:

Table 1. Input dataset.

stu_ID abcd-AA	Result	Science	commerce	civics	english	hindi
45	-0.054829956	26	29	33	26	41
44	0.072182266	22	39	34	17	37
41	-0.242498969	17	29	32	31	22
34	0.119176859	26	44	33	14	39
31	-0.163328092	34	33	31	18	31
25	-0.181906429	23	24	25	22	28
24	-0.197937552	26	33	36	19	37
22	-0.102977317	32	32	32	23	33
13	0.187188343	20	33	30	16	41

B. Data Mining Method

Input data is pre-processed using WEKA filters, here unused data is removed. The preprocess panel is shown in Fig 2. On this pre-processed data, regression is applied in order to derive the relation between independent and dependent variables. Independent variables are considered as the input values to the model and dependent variable is considered as output of the model. Regression finds the relationship between dependent and independent variables. In the dataset we considered, the dependent variable is "result" and subjects are the independent variables. Here we establish the relationship between the overall results of students to their performance in individual subjects.

Here, linear regression is used. The data we considered is linear in nature. Regression is given by the formula:

$$Y=a+bX$$

Where,

Y = Dependent variable
a= Intercept
b= Slope
X= Independent variable

C. Output

The output of the data pre-process using WEKA is shown in Fig 3.

Fig. 2. Image dipecting primary preprocessed results.

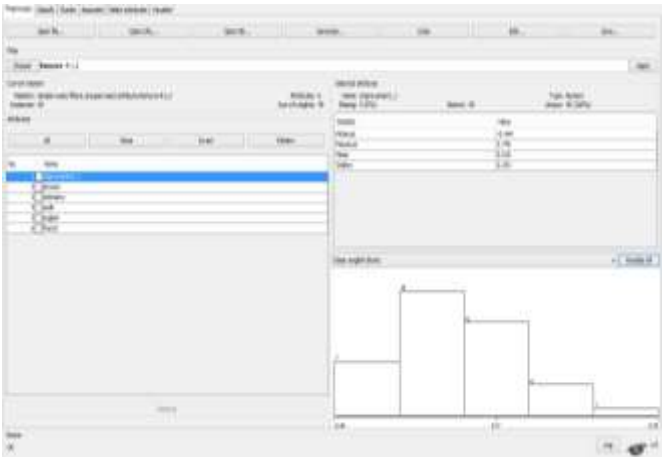

Table 2. Output dataset.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	0.13983	0.48598	0.29	0.7753
Science	1	0.01008	0.00872	1.16	0.2553
Civics	1	-0.00658	0.00632	-1.04	0.3047
Commerce	1	-0.00281	0.01069	-0.26	0.7943
English	1	-0.01578	0.00640	-2.47	0.0189

4. TRIPLET EXTRACTION

Stanford CoreNLP provides a set of natural language analysis tools which can take raw text input and give the base forms of words, their parts of speech, whether they are names of companies, people, etc [6][7]., normalize dates, times, and numeric quantities, and mark up the structure of sentences in

terms of phrases and word dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, etc.

POSTaggerAnnotator class generates parts of speech annotation. Labels tokens with their POS tag.

D. Input Files

1. Total significance of results is 3.
2. R-Square represents fitness of the model.
3. The model is fitted by 27%.
4. English has the major affect on the Result.
5. The parameter estimate of Model is negative.

Loading POS Tagger model ... done (Total significance of results is 3. 2.553s)

Output:Total_JJ significance_NN of_IN results_NNS is_VBZ 3._VBG
Array list s3 is:[Total_JJ, significance_NN, of_IN, results_NNS]
Array list s4 is:[is_VBZ, 3._VBG]
Subject'0'==========significance
Subject'1'==========results
Predicate ===========[3.]
object is ========= []

R-Square represents fitness of the model
output:R-Square_DT represents_VBZ fitness_NN of_IN the_DT model_NN
Array list s3 is:[R-Square_DT]
Array list s4 is:[represents_VBZ, fitness_NN, of_IN, the_DT, model_NN]
Predicate ===========[represents]
object is ========= [model]

The model is fitted by 27%
output:The_DT model_NN is_VBZ fitted_VBN by_IN 27%_CD
Array list s3 is:[The_DT, model_NN]
Array list s4 is:[is_VBZ, fitted_VBN, by_IN, 27%_CD]
Subject'0'==========model
Predicate ===========[fitted]
object is ========= []

English has the major affect on the Result
output:English_NNP has_VBZ the_DT major_JJ affect_VBP on_IN the_DT Result_NN
Array list s3 is:[English_NNP]
Array list s4 is:[has_VBZ, the_DT, major_JJ, affect_VBP, on_IN, the_DT, Result_NN]
Subject'0'==========English
Predicate ===========[affect]
object is ========= [Result]

The parametr estimate of model is negative
output:The_DT parametr_NN estimate_NN of_IN model_NN is_VBZ negative_JJ

Array list s3 is: [The_DT, parametr_NN, estimate_NN, of_IN, model_NN]

Array list s4 is: [is_VBZ, negative_JJ]
Subject'0'==========parametr
Subject'1'==========estimate
Subject'2'==========model
Predicate ===========[is]
object is ========= []

II. RESOURCE DESCRIPTION FORMAT

After Apache Jena API uses Java system for RDF providing support for manipulating RDF models, parsing RDF/XML.

Generated RDF for Triples

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:SASResult="http://sasresults.edu#" >
 <rdf:Description rdf:about="http://sasresults.edu#results">
  <SASResult:Subject>'results'</SASResult:Subject>
  <SASResult:Predicate>'Total'</SASResult:Predicate>
  <SASResult:Object>'null'</SASResult:Object>
</rdf:Description>
 <rdf:Description rdf:about="http://sasresults.edu#model">
  <SASResult:Subject>'model'</SASResult:Subject>
  <SASResult:Predicate>'null'</SASResult:Predicate>
  <SASResult:Object>'null'</SASResult:Object>
</rdf:Description>
 <rdf:Description rdf:about="http://sasresults.edu#model">
  <SASResult:Subject>'model'</SASResult:Subject>
  <SASResult:Predicate>'null'</SASResult:Predicate>
  <SASResult:Object>'fitted'</SASResult:Object>
</rdf:Description>
 <rdf:Description rdf:about="http://sasresults.edu#Result">
  <SASResult:Subject>'Result'</SASResult:Subject>
  <SASResult:Predicate>'major'</SASResult:Predicate>
  <SASResult:Object>'null'</SASResult:Object>
 </rdf:Description>
  <rdf:Description rdf:about="http://sasresults.edu#eng">
  <SASResult:Subject>'eng'</SASResult:Subject>
  <SASResult:Predicate>'negative'</SASResult:Predicate>
  <SASResult:Object>'null'</SASResult:Object>
 </rdf:Description>
</rdf:RDF>
```

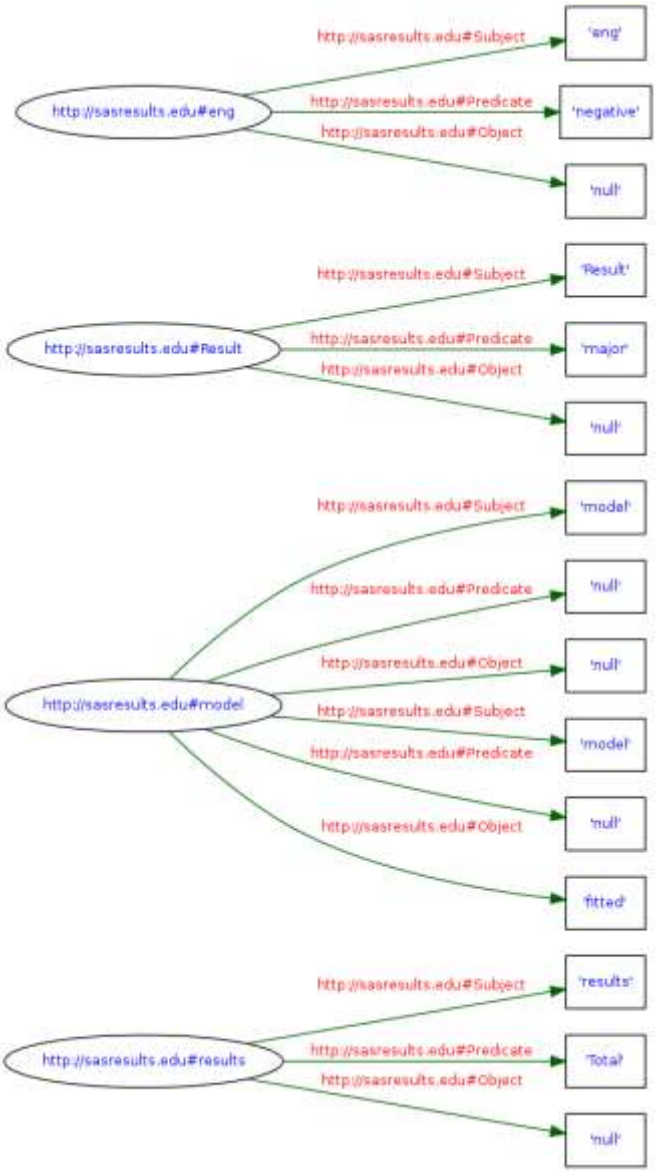

Fig. 3. RDF graph model .

5. CONCLUSION

In this paper, we have proposed a new way to generate and present knowledge from large amounts of potentially heterogeneous and distributed data set. Resulted RDF aims at describing and formalizing entities from the domain of data mining and knowledge discovery. This system will help to build expert automated decision support system based on the data mining results.

6. REFRENCES

[1] Li Ding, Pranam Kolari, Zhongli Ding, and Sasikanth Avancha, Using Ontologies in the Semantic Web: A Survey, book-chapter in Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems, 2006.

[2] M. Taye , "Understanding Semantic Web and Ontologies: Theory and Applications", J. of Computing, Vol. 2(6), June 2010, NY, USA, ISSN 2151-9617.

[3] R. Jasper and M. Uschold. A framework for understanding and classifying ontology applications. In Proceedings of the IJCAI99 Workshop on Ontologies and Problem-Solving Methods(KRR5), 1999.

[4] Mizoguchi, R.: Tutorial on ontological engineering - part 3: Advanced course of ontological engineering. New Generation Comput 22(2) (2004)

[5] Smith, B.: Ontology. In: Blackwell Guide to the Philosophy of Computing and Information, pp. 155–166. Oxford Blackwell, Malden (2003).

[6] Gruber, T.R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5, 199-220.

[7] Ganter, B.; Stumme, G.; Wille, R. (Eds.) (2005). Formal Concept Analysis: Foundations and Applications. Lecture Notes in Artificial Intelligence, no.3626, Springer-Verlag. ISBN 3-540-27891-5.

[8] Anuj Tiwari, Dr. Kamal Jain (2014), “Ontology Driven Architecture for Web GIS”, India Geospatial Forum 2014, February 5–7, 2014, 60, Hyderabad, India.

International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 11 - 17, 2015, ISSN:- 2319–8656

Proposing a scheduling algorithm to balance the time and cost using a genetic algorithm

Ali Akbar Faraj
Department of Computer
Science and Research Branch
Islamic Azad University Kish, Iran

Ali Haroon Abadi
Member of Science Board of
Computer Group in Azad Islamic
University of Tehran Center

Abstract: Grid computing is a hardware and software infrastructure and provides affordable, sustainable, and reliable access. Its aim is to create a supercomputer using free resources. One of the challenges to the Grid computing is scheduling problem which is regarded as a tough issue. Since scheduling problem is a non-deterministic issue in the Grid, deterministic algorithms cannot be used to improve scheduling.
In this paper, a combination of genetic algorithms and binary gravitational attraction is used for scheduling problem solving, where the reduction in the duty performance timing and cost-effective use of simultaneous resources are investigated. In this case, the user determines the execution time parameter and cost-effective use of resources. In this algorithm, a new approach that has led to a balanced load of resources is used in the selection of resources. Experimental results reveals that our proposed algorithm in terms of cost-time and selection of the best resource has reached better results than other algorithm.

Keywords: Grid computing, Static timing strategies, Genetic algorithms, Local search algorithm following the binary gravitational attraction, Optimizing Time

1. INTRODUCTION

Grid is a form of distributed computing systems which is available on a network and appears to the user as a large virtual computing system [4]. One of the most important parameters in the Grid environment is grid scheduling and load balancing services. Therefore, a correct and reliable scheduler is needed to increase the efficiency of the grid. Unfortunately, the dynamic nature of the Grid resources and users' demands has led to complexity of Grid scheduling problem and optimized assigning of tasks to resources. This would have prompted researchers to use innovative algorithms for solving this challenge. A scheduler in grid environment should provide a time and cost effective scheduler for users through using users' parameters, without engaging users in the complexity of such an environment. Genetic algorithm is one of the best innovative algorithms due to the fact that it generally investigates the problem from several different directions at once. Therefore, genetic algorithm is used to solve many optimization problems [7]. In this paper, after reviewing the strengths and weaknesses of previous approaches, the combination of genetic algorithm and local search following the gravitational force will be examined with an aim of reducing the simultaneous implementation of tasks time and costs. In this algorithm, a new approach has been used in the selection of resources.

The following sections of this paper are as follows: Section 2 describes the literature. Section 3 introduces the related work on scheduling and Section 4 is dedicated to the algorithm proposed. Performance evaluation of the proposed algorithm is presented in Section 5. Section 6 presents the conclusions.

2. BACKGROUND

2.1 Scheduling Method

The scheduling task issue is considered as a tough challenge which is composed of n tasks and m resources. Each task must be processed by a machine and does not stop until the end of the performance. We used ETC matrix model described in [2]. The system assumes that the expected execution time for each task i, on every resource j is predetermined and is located in the matrix ETC, ETC [i, j]. Here, makespan is regarded as the maximum completion time in Completion-Time [i, j], calculated in the following equation (1)[10]:

Makespan=Max(Completion- Time[i,j]1≤i≤N,1≤j≤M (1)

In the above equation, Completion-Time [i, j] is equal to thetime when the task i on the source j is completed and it is calculated in equation (2):

Completion-Time[i,j]=Ready[M]+ETC[i,j] (2)

2.2 Genetic Algorithm

Inspired by the evolution of organisms in nature, genetic algorithm is optimizing problems. Based on the evolution of living organisms, species can survive in nature. Genetic algorithm randomly selects chromosomes. Combination is a process that displaces determined sequence of the chromosomes. Mutation is a process that changes the determined sequences of chromosomes with multiple mapping functions, new to the current population. Combination and mutation operations are done randomly. After this operation, a new population is generated, then the population will be evaluated and the process will be repeated several times until the accomplished process criteria are authenticated [5].

2.3 Local search algorithms following binary gravitational attraction

Voudouris, et al. proposed GELS algorithm for the first time in 1995 to surf in a search and tough problem solving space [11]. In 2007, Webster offered it as a powerful algorithm and called it the Gravitational Attraction Algorithm. This algorithm is based on the principle of gravity force and imitates this natural process to surf within a search space. It is assumes that there are only the gravity and motion laws governing. Each response has different neighbors which can be classified based on a criterion related to the problem. Obtained neighbors in each group are called neighbors in that dimension. Each dimension has got an initial velocity. After an initial velocity is defined for each dimension, the greater velocity the dimension has, the better response will be provided to the problem. In each dimension, the neighbors responses to simmilar methods are obtained from the current response. It is so-called obtaining neighbor's response from the current response in a specified dimension. For each dimension of response, an initial velocity is defined. The greater the velocity is, the more appropriate response will be provided for the issue [5]. The mass of an object is determined based on its fitness, according to equation (4). The fiti (t) is the the suitability of mass i, worst (t) is the worst fitness at time t, and best (t) is the best fitness at time t.

$$M_{ii}(t) = \frac{M_i(t)}{\sum_{j=1}^{N} M_j(t)} \quad (3)$$

$$M_i(t) = \frac{fiti(t) - worst(t)}{best(t) - worst(t)} \quad (4)$$

The mass of dimension d is accelerated in a way that is proportional to the force exerted on the mass in that (Fid) direction divided on the inertial mass or inertia (Mii), expressed in equation (5):

$$a_i^d(t) = \frac{F_i^d(t)}{M_{ii}(t)} \quad (5)$$

Elitist gravitational search algorithm is introduced for solving continuous problems in search space. In fact, Vid has the likelihood of zero or one of Xid instead of indicating the speed in binary version. Its formula uses equation (6) to update the position of each dimension [10]

$$X_i^d(i,j) = \begin{cases} \sim X_i^d(t) & if\ rand < S(V_i^d(t+1)) \\ X_i^d(t) & if\ rand \geq (V_i^d(t+1)) \end{cases} \quad (6)$$

Each mass velocity is the sum of the ratio of its current mass speed and mass acceleration, updated according to equation (7). It should be noted that because of determination of the range for the initial velocity, if adding momentum to an entry of the initial velocity vector cause its value to exceed from the specified range, its value would be set in this range [9].

$$V_i^d(t+1) = rand_i \times V_i^d(t) + a_i^d(t) \quad (7)$$

3. Previous Research

Depending on the size and dynamics of the Grid, deterministic algorithms cannot be useful for solving the scheduling problems. This has led researchers to research innovative algorithms for these problems.

A valuable technique for the design of effective scheduling is proposed using genetic algorithm and aims to minimize the response time tasks [12]. The results show that the scheduler has a very high speed and decreases time response. Grid users pay the costs to holders of financial resources for the sake of using such resources [1]. This brings about an incentive for resource owners to share their resources. This model leads to significant economic costs and benefits for users and owners of the resources. As a result, the costs and economic benefits are considered in the objective function of scheduler in most of the scheduling algorithms [11]. A new algorithm has been suggested by combining genetic algorithms and gravitational attraction to solve the scheduling task problems in grid environment. The cost of using this algorithm is ignored [8].

In reference [5], a genetic algorithm, in which disorganized variables are used instead of random variables to produce chromosomes, is proposed. This makes the solution in the search space spread, prevents the algorithm premature convergence, and brings about better solutions and product in a shorter time. This paper uses an algorithm for scheduling tasks in grid gravitational attraction [3].

Most of these procedures have considered completion time and a few other implementation costs. However, they failed to establish a balance and balanced load among them. Thus, in Section 4, a new method will be discussed to investigate the balance and balanced load in a Grid environment.

4. . THE PROPOSED ALGORITHM

The proposed scheduling uses the combination of genetic algorithms and binary gravitational attraction for scheduling tasks in grid. Since the performance of genetic algorithms depends largely on how chromosomes are encoded, we will further investigate it. In this way, a chromosome shows a solution. Each solution in this algorithm is encoded as a row of natural numbers matrix. For example, we have assumed a set of n tasks T = {T1, T2, T3, ..., Tn} and a set of m resource P = {P1, P2, P3, ..., Pm}. In this method, each gene represents a cell as a function and the chromosomes' lengths are considered as the number of input functions. The contents of each cell of the matrix is as the resource and its value in chromosomes can be random between 1 and m. Figure 1 shows an example of the chromosomes encoding.

Figure 1. The chromosomes encoding in BGAGSA algorithm

4.1 The objectives and fitness function in the proposed algorithm

The first objective of the proposed algorithm is to reduce the longest completion task time of all processors in the system with makespan. In the proposed method, the user is allowed to enter the maximum completion time and maximum costs of task performance. Here suppose that Lei represents the task i and R is the processing speed of the resource. The execution time of task i on j resource can be calculated by the equation (8).

$$T(i,j) = \frac{Le_i}{R_j} \quad (8)$$

The completion time of task i on resource j can be calculated based on equation (9).

$$CompleteTime(i,j) = T(i,j) + TransferT + wait(i,j) \quad (9)$$

In the above equation, TransferT and wait (i, j) are respectively the data transfer time and waiting execution time. Thus, the makespan system can be calculated using equation (10), where Aj is a set of tasks allocated to resource j.

$$CompleteTime(j) = \frac{(\sum_{k \in A_j} CompleteTime)}{R_j} \quad (10)$$

$$F_{Time} = Max\{CompleteTime(j)\}\ 1 \le J \le M \quad (11)$$

The second objective of our algorithm is the total cost which must be minimized. Suppose that Rj representing a fixed unit price for resource j and is considered fixed and TransferC is the cost data transfer. The cost of task i on resource j would be calculated by equation (12).

$$cost(i,j) = \big(T(i,j) \times P_j\big) + TransferC \quad (12)$$

The cost of task i accomplishment on resource j can be computed using equation (13).

$$Cost(j) = completeTime(j) \times P_j \quad (13)$$

Therefore, the overall cost of a solution (chromosome), which represents the cost associated with a chromosome in the population, can be expressed using equation (14).

$$F_{cost} = \sum_{J=1}^{M} Cost(j) \quad (14)$$

4.2 First fitness function

When offering tasks, some users are concerned with the task cost and others with completion time. Accordingly, the user can enter the values of cost and time (XT and XC) in the proposed method. These values are placed in the range [0,1] and sum of them is equal one. For example, if Xc is equal to .7, this means that the user is 70% concerned with costs and 30% concerned with completion task time. The scheduler should find a resource for performing tasks and that resource should improve costs up to 70% and time up to 30%. The most important task scheduling objective is to minimize the time and cost of completing tasks. Given the above definitions, a chromosome fitness function can be calculated in equation (15):

$$Fitness = X_t \times \frac{1}{F_{Time}} + X_c \times \frac{1}{F_{Cost}} \quad (15)$$

Where FTime and FCost are the longest completion time of tasks and the overall cost of a solution. As the equation shows, the less the values of FTime and FCost are, the more the value of the fitness function will be and shows that the more suitable solution for the scheduling problem it is. In the proposed algorithm, to select chromosomes (parents) from competition operator and intersection operations from partially mapped crossover operator in mutation phase after selecting a chromosome in previous stage, a gene is randomly selected and its source field value randomly changes between 1 and m. The main purpose of these mutations is changing the resource parameter of a function in order for a task to be applied on better processing resources.

4.3 Second fitting Function

Since the main objective of the proposed scheduling algorithm is to minimize the total length of time and reduce the cost of implementing tasks, several chromosomes are found and their total scheduling length is the same. Thus, a chromosome with more balanced load is selected in the second fitness function. To calculate the balanced load, we calculate the sum of the standard deviation of productivity. As a result, we must calculate the efficiency of resources. Productivity is calculated in equation (16). We initially calculate the greatest performance time.

$$\boldsymbol{U = Max\{T_{ij} + t_{ij}\}}$$
(16)

In the above equation, Tij is the sub-task processing time and t is data transmission time. They are calculated to obtain the greatest performance time to be put into the variable U. If the efficiency of resource i in the execution of the task j is shown by E the relationship, it can be calculated in equation (17).

$$\boldsymbol{E_i = \frac{T_{ij}+t_{ij}}{U}}$$
(17)

In the amount of U is then calculated using equation (16). According to equation (17), the efficiency of each source will be obtained. To obtain the standard deviation, we calculate the mean total productivity using equation (18).

$$\boldsymbol{E = \frac{\sum_{i=1}^{m} E_i}{m}}$$
(18)

To do this, we need to add all the productivity of resources together and then divide it by the number of sources. Then, we calculate the load balance using equation (18) at which the efficiency of all resources is computed. Finally, if we have μ2 = 0, the maximum load reduction will be obtained.

4.4 Local search algorithm parameters imitating the binary gravitational attraction

At this stage, because the genetic local search algorithm is weak, some solutions, which are similar in terms of overall length of scheduler, are given to the GELS algorithm to have a solution to their neighbors. Each current solution has different neighbors. Each of them is obtained based on a change in the current solution, that is, shift towards a neighbor solution [4]. In the proposed method, a solution represents a chromosome and solution dimensions are considered as each gene from chromosome (Dimensions of the solution are the neighbor solutions which are obtained through changing the current solution). To determine the representation of solutions with the aim of finding an appropriate mapping, it is performed between solutions and GELS factors. Solutions are encoded in the form of m × n where m represents the source and n represents the task and the element XK (i, j) indicates that task j is assigned to an agent k in the resource i. For example, tasks 4 and 5 are executed simultaneously in a resource. Each task can only be assigned to a single resource. Each resource can have multiple tasks in sequential form. The encryption is performed using equation (19) to update the position of each dimension [13].

$$\boldsymbol{X_k(i,j) = \begin{cases} 1 & \forall_k(i,j) = Max\{V_k(q,j)\} \\ & \forall_q \in \{1,2,\dots\dots\dots\dots,m\} \\ 0 & otherwise \end{cases}}$$
(19)

In the above equation, a resource is dedicated to a task Xk and the relevant element in Vk is not assigned to any other resource. The number of entries of initial velocity vector has been considered equal to the number of solutions. The initial velocity vector entries are initially assigned random numbers between 1 and maximum initial velocity [5]. The search space is considered as a set of N mass. The position of each mass in the search space is a point of the space that is a solution to the problem. Neighbor solution of each current solution is equal to a solution where the source assigned to task changes. This is done in a way that an initial velocity is given to each dimension of the solution. This work is done randomly and its value is between one and maximum speed. The gene with the highest speed is selected and its value will change randomly (with a number between 1 and M) [5].

After obtaining the solution of the current neighbor, the fitness value of the solution of neighbor is calculated using the equation (15). If the neighboring solution is improved in comparison with the current solution, it replaces its parent chromosome in the new population. Otherwise, it will not be copied into the new population, and its value is maintained to calculate the mass of each particle. Then, the value of acceleration is calculated between the neighboring solution and the current solution, and its value is added to the initial velocity vector, related to the dimension from which the neighboring solution is obtained, in order for the initial velocity vector to be updated. It should be noted that due to the determination of the range for the initial velocity, if by adding the gravitational force to an entry in the initial velocity vector, its value exceeds the specified range, its value is set in this range interval. Given the above, the mass of neighboring solutions is calculated through equation (20) [13]

$$\boldsymbol{M_k = \frac{f(X_k)-worst}{\sum_{i=1}^{N} f(X_i)-N\times worst}\, k = 1,2,\dots\dots}$$
(21)

Where F(Xk) is the fitness value of particle, and worst is the worst value among all the particles. The acceleration matrix of K factors is calculated through equation (22) [13].

$$\boldsymbol{a_k(i,j) = \sum_{j\in K_{best}} G \times \frac{rand_q M_q}{R+\varepsilon}(X_q(i,j) - X_k(i,j))}$$
(22)

According to the above equation, the mass of i with an acceleration equal to a is drawn to the mass of j (ak (i, j)), in

which Kbest is a collection of first K factor with the best performance value and the greatest mass. Mq is the mass of Q factor and G is the gravitational constant. The pseudo-code of the proposed algorithm is presented in figure 2.

step1: initialization
step2: 2.1. Generate the k number of random Chromosome with lenght n
2.2. Speed_vector[1..n] = The initial velocity for each dimension
2.3. Location_vector[1..n] = The value of the initial position of each particle
2.4. Setting the Maximum: Maximum-Time (Mt) and Maximum-cost (Mc) according to the user's requirement.
step3: Compute Makespan and cost for all chromosomes.
step4: repeat
4.1. Evaluate all individuals using formula 15
4.2. Select the P/2 members of the combined population based on minimum fitness, to make the population the next generation.
4.3. Crossover & Mutation
step5: The best chromosomes from the genetic algorithm as a solution to it current GELS will be producing a neighboring chromosomes.
step6:
6.1. direction = max(speed_vector[...])
6.2. change the velocity_vector[index] of current_solution with random integer between 1 and Max Velocity.
6.3. Fitness of the neighboring chromosomes by equation (15) is calculated and and saving the worst fitness
6.4. if direct chromosomes < neighbor chromosome Neighbor chromosome is selected as the best solution
setp7:
7.1. Calculated the mass of each particle with using formula (22)
7.2. Calclulate Acceleration Matrix for 'k' factors with using formula (21)
step8:
8.1. Update the factors by using (19)
8.2. Update Velocity_Vector for each dimension by acceleration matrix of chromosome with using formula (7)
until a terminating criterion is satisfied

Figure 2.Pseudocode the algorithm proposed

5. DISCUSSION

The results of the evaluation of the proposed algorithm with the two algorithms of [6] GA and [7] Genetic-Variable neighborhood search for scheduling independent tasks are presented in this section. All experiments have been done on a system running Windows XP operating system with configuration of 3 GHz CPU and 4GB of RAM. Table (1) indicates the amounts of parameters of genetic and binary gravitational emulation local search algorithms.

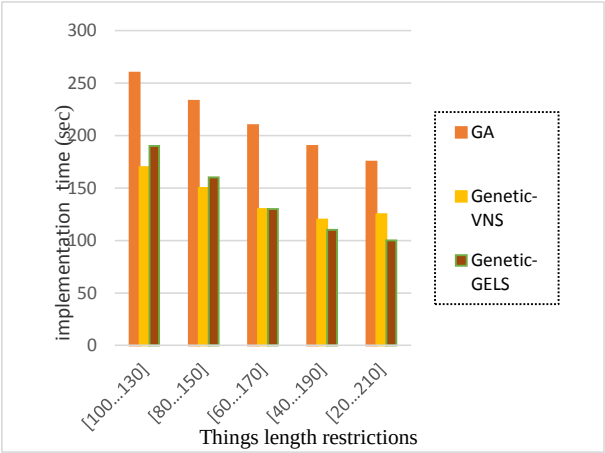

Figure 3. Comparison of makespan of algorithms

GA	Number of generations	100
	Selection	Contesting
	Combined rate	0.85
	Mutation rate	0.02
GSA	Initial position	Between 1 and n
	Initial velocity	Between 1 and the maximum speed
	Maximum speed	The size of the input functions
	Neighborhood radius (R)	1
	Gravitational constant (G)	6.672

To evaluate the proposed algorithm, a series of simulations was considered on 10 sources based on the change in the number of users' tasks and payable budgets by the customers. In the first experiment, aimed at optimizing the time, we have experimented the proposed algorithm with two other scheduling algorithms through applying a change in the number of tasks and a fixed budget. The number of 20 iterations have been performed in this algorithm in order to obtain the execution time of task using existing algorithms, as the results have been uncertain and the results of each performance can be different to some extent from the previous one. The results of the first experiment (figure 3) indicate that the proposed algorithm has a higher performance than the other two algorithms in terms of task completion time. Furthermore, by increasing the number of iterations, the overall performance time increases in all algorithms.

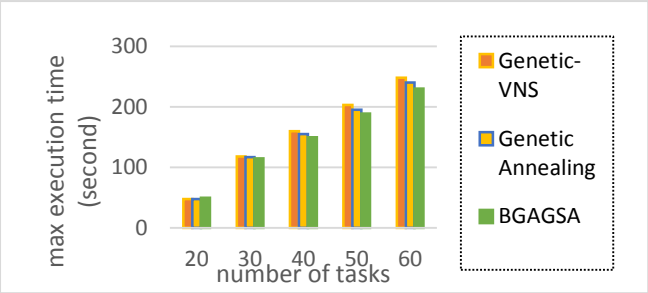

Figure 4. Results of time-optimization algorithms with different budgets

Having the purpose of optimizing time, the second experiment is compared through providing different budgets. The proposed algorithm is evaluated along with several other scheduling algorithms. The budget parameter has been considered as the variable. As it can be seen in figure 4, Genetic-GELS algorithm functioned better than other algorithms in all cases, and by increasing the budget, the amount of execution time is reduced. In fact, the more specified budget set by the user, the less time is required for algorithms to perform practical tasks.

In the third experiment, the effect of heterogeneity of tasks on the performance of the proposed algorithm and other targeted algorithms was investigated. The length of the tasks is considered variable. As it can be seen in figure 5, the time changes differently due to an increase in the heterogeneity of the tasks in each of the algorithms. A good algorithm is expected to be able to make use of the heterogeneity of tasks and reduce the execution time of tasks. As it is shown clearly in the figure, the proposed algorithm has achieved a better performance compared to other algorithms.

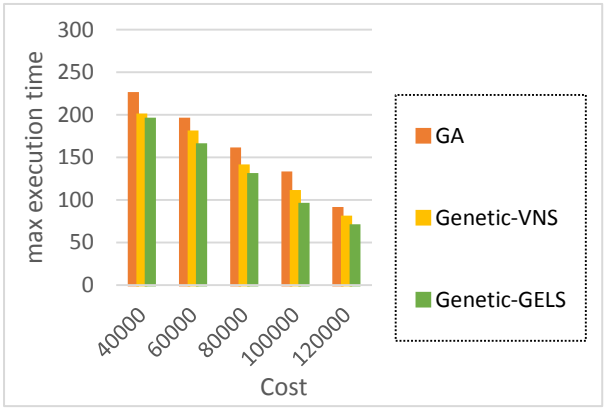

Figure 5. Results of time-optimization algorithms with different heterogeneities

7. REFERENCES

[1]Buyya,R., Giddy J., Abramson,D., A case for economy grid architecture for service-oriented grid computing , in: 10th IEEE Internat. Heterogeneous Computing Workshop , San Francisco, CA, April 2001.

[2]Braun, T. D., Siegel, H. J , A taxonomy for describing matching and scheduling heuristics for mixed machine heterogeneous computing systems, Proceedings of the 17th IEEE Symposium on Reliable Distributed Systems, pp. 330-335, 1998.

[3]Cruz-Chavez, M., Rodriguez-Leon, A., Avila-Melgar, E., Juarez-Perez, F., Cruz-Rosales, M. and Rivera-Lopez, R,"Genetic-Annealing Algorithm in Grid Environment for Scheduling Problems, Security-Enriched Urban Computing and Smart Grid Communications in Computer and Information Science", Springer, Vol. 78, pp. 1-9, 2010.

[5]Ghaedrahmati V.,Enayatallah Alavi S.,Attarzadeh I.,"A Reliable and Hybrid Scheduling Algorithm based on Cost and Time Balancing for Computational Grid "ACSIJ Advances in Computer Science: an International Journal, Vol. 3, Issue 3, No.9 , May , ISSN : 2322-5157, 2014.

[6]Gharooni fard. G, Moein darbari. F, Deldari. H, Morvaridi. A, "Scheduling of scientific workflows using a chaos- genetic algorithm", International Conference on Computational Science ICCS, pp. 1439- 1448, 2010.

[7]Holland, j. "Adaptation in Natural and Artificial Systems", AnnArbor,MI: Univ. of Michigan Press. 220, no. 4598, pp. 671- 680. 1975.

6. CONCLUSIONS

In the proposed method, the researchers have combined genetic algorithm with binary gravitational emulation local search algorithm which has been used for scheduling in grid environment. The proposed algorithm, in which a weighted objective function is used considering the degree of importance of time and cost of user's projects, gives more freedom for specifying the time and cost of users' projects. In the use of this algorithm, similar to the objective function, the time and the cost, along with their weight, are considered based on the user's perspective. The purposes of the proposed scheduling are to minimize completion time and cost of implementation of tasks for different tasks of users simultaneously.

The proposed scheduling is compared with two algorithms of [6] GA and [7] Genetic-Variable neighborhood search with regard to the parameters of time and cost. The results are investigated based on cost and user requests in different charts. They show that if we have limited number of resources as well as high number of duties, the proposed scheduler has the best performance compared to the other three schedulers in reducing the overall time and cost of scheduling. The results of experiments show that the hybrid genetic algorithm and the gravitational attraction can reach a high performance regarding creation of a balance between cost and tasks implementation scheduling. Further research can be conducted with regard to the practice of resource allocation to works in the proposed algorithm through using an approach based on fuzzy logic and using of fuzzy inference system.

[8]Kardani-Moghaddam, S., Khodadadi, F., Entezari-Maleki, R., and Movaghar,A., "A Hybrid Genetic Algorithm and Variable Neighborhood Search for Task Scheduling Problem in Grid Environment", Procedia Engineering, 29, pp. 3808-3814,2012.

[9]Pooranian, Z., Harounabadi, A., Shojafar, M., Hedayat,.N, "New hybrid algorithm for task scheduling in grid computing to decrease missed task", World Acad Sci Eng Technol 55,924–928,2011.

[10]Rashedi, E., Nezamabadi-Pour, H., Saryazdi, S.,"BGSA: binary gravitational search algorithm". Nat. Comput. 9(3), 727–745, 2010.

[11]Shojafar, M., Pooranian, Z., Abwajy,J.H., Meybodi, M., ,"An Efficient Scheduling Method for Grid Systems Based on a Hierarchical Stochastic Petri Net" ,Journal of Computing Science and Engineering,Vol. 7, No. 1, March, pp. 44-52,2013.

[13]Young,L., McGough,S., Newhouse,S. and Darlington,J., Scheduling Architecture and Algorithms within the ICENI Grid Middleware, in Proc. of UK e-Science All Hands Meeting, pp. 5-12,

Nottingham, UK, September 2003.

[14]Zhang, L., Chen, Y., Sun, R., Jing, S., Yang, B., "A Task Scheduling Algorithm Based on PSO for Grid Computing", International Journal of Computational Intelligence Research, Vol 14. , No.1 , pp 37-43,2008.

[15]Zarrabi A.,Samsudin K.,"Task scheduling on computational Grids using Gravitational Search Algorithm",Department of Computer and Communication Systems Engineering,University Putra Malaysia,Volume 17,issue 3, December,ISSN: 1573-7543,pp 1001-1011. 2013.

International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 18 - 23, 2015, ISSN:- 2319–8656

Cluster as a Service (CaaS) in Secure Deduplication System

R.Nalla Kumar
Department of CSE
Regional Centre
Anna University
Coimbatore.India

A.Kumari savitha sree
Department of CSE
Regional Centre
Anna University
Coimbatore, India

X.Alphonseinbaraj
Department of CSE
Regional Centre
Anna University,
Coimbatore, India

Abstract: Data deduplication is one of compression method of data for eliminating duplicate copies of repeating data that is mainly used to reduced the amount of storage space and bandwidth. Cloud computing systems have been made possible through the use of large –scale clusters,service –oriented architecture (SOA),web services ad virtualization.While the idea offering resources via web services is common place in cloud computing ,little attention has been paid to clients themselves specifically ,human operators.Despite that clouds host a variety of resources which in turn are accessible to variety of clients ,support for human users is minimal .To provide better security ,this paper makes the first attempt to formally address the problem of authentication ,integrity and availability.By using Tag generation , moreover one of additional cloud storage service such that Cluster as a Service (CaaS) can make secure deduplication possible and reduced cloud storage space.With out key generation ,attribute based encryption makes secure data deduplication in the cluster of computers.

Keywords: Deduplication,PoW,Cluster as a Service ,Identification Protocol

1 INTRODUCTION

Cloud computing provide unlimited **" virtualized resources "**to users as services across the whole Internet ,while hiding platform and implementation today. Cloud Infrastructure providers are establishing cloud centers to host a variety of ICT services and platforms of world wide individuals ,innovators and institutions. Cloud Service Providers(CSP) are very aggressive in experimenting and embracing the cool cloud ideas and today every business and technical services are being hosted in cloud to be delivered to global customers ,clients and consumers over the Internet communication infrastructure .For example Security as a service (SaaS) is a prominent cloud –hosted security service that can be subscribed by spectrum of users of any connected device and users just pay for the exact amount or time of usage .Besides the modernization of legacy applications and positing the updated and upgraded in clouds ,fresh applications are being implemented and deployed on clouds to be delivered to millions of global users simultaneously affordably.While web services have simplified resource access and management ,it is not possible to know if the resource(s) behind the webservice is (are) ready for request.Clients need to exchange numerous message with required Web services to learn the current activity of resources and thus face significant overhead loss if most of the web services prove ineffective.Furthermore ,even in ideal circumstances where all resources behind Web services are the best choice,clients still have to locate the services themselves . Finally ,the Web services have to be stateful so they are able to best reflect the current state of their resources.

Although data deduplication provide lots of benefits and advantages ,security and privacy concerns sensitive data are susceptible to both insider and outsider attacks . Normal traditional deduplication is incompatible with encryption. Specifically different users produces different ciphertext makes data deduplication ineffective and not feasible .Convergent encryption [1],[2] has been proposed to enforce data confidentiality while data deduplication is feasible .

It encrypt and decrypts a data copy with *convergent key* and further to avoid unauthorized entry in system, a secure proof of ownership[5] is also needed to provide the proof that the user indeed owns the same file while duplicate is found . In additionally tag was generated by attribute based encryption .

This method is different from traditional techniques . In traditional method,each time user can access the file with their own private key ,but here attribute based encryption done such that generating tag and by which privileges[6][4] is associated with that. Each file is uploaded to the cloud is also bounded by set of privileges to specify which kind of users is allow to perform the duplicate check and access the file.The user can find the duplicating of the file if it is stored in cloud .In our system, we need to consider the three things as follows. 1. Cloud Management System 2. Virtual cluster administrations 3.user .

1.1 CONTRIBUTIONS

In this paper,

1) Performing convergent encryption with differential privileges and tag generation which is used to avoid duplicate without generate key in client /user side.

2) The users without corresponding privileges cannot perform the duplicate check .For example ,in a

company , many privileges will be assigned to employees.In otherwords ,no differential privileges have been considered in the deduplication based on convergent encryption techniques in traditional deduplication methods.

3) Introduce the first provably –secure deduplication method in Cluster based Service

1.2 Organizations

The rest of this paper proceeds as follows. Section 2 briefly revisit some preliminaries of this paper. Section 3 propose the system model for secure data protection. In section 4, implementation progress of the secure duplication system is described. In section 5, some other related work regarding this system is described.Some experimental result are shown in the section 6. In section 7, some future work and some other ideas to enhance security is described. Finally conclusions are drawn in section 8.

2 PRELIMINARIES

In this section, we are going to consider attribute based encryption and review some secure primitives used in our secure deduplication

2.1 Cluster as a Service (CaaS) implementation:

The main role of CaaS is to (i) provide easy and intuitive file transfer so clients can upload file(s),(ii) offer an easy to use interface for clients to monitor their process.The CaaS does this by allowing clients to upload files as they would any web page while carrying out the required data transfer to the cluster transparently.

By hiding hardware and software features of cluster ,the CaaS provides higher level abstraction. Clients only receive the minimal amount of the data and provide the web pages to deploy,run and control execution of process.Because clients to the cluster cannot know how the data storage is managed,the CaaS offes a simple transfer interface to clients while addressing the transfer specified.

In some other experiments,CaaS was implemented by Windows Communication Foundation of .Net using web services .The problem is client(s) exchanges the numerous messages in case of activation of resources.

2.2 Proof of Ownership(PoW):

Halevi et al [10] proposed the notion of "Proof of Ownership(PoW)" for deduplication system.Such that client prove their authentication without uploading their files .Several PoW constructions based on the Merkle-Hash Tree proposed [10] to enable client side deduplication which include the bounded leakage setting and another schemes such that Pietro and Sorniotti[7] proposed some bit positions based file proof.Now recently Ng et al [8] proposed some PoW method but that not address that how to minimize the key management overhead etc.,

In our system PoW is used to enable users to prove their ownership in order to found deduplication occur in system.Generation of Tag is mainly used to detect the duplication .Virtual machine technology makes it very flexible and easy to mange resources in cloud computing environments,because they improve the utilization of such resources by multiplexing many virtual machines on one physical host(server consolidation).These machines can be scaled up and down on demand with a high level of resources abstraction .

2.3 Proof & Verify protocol:

There are identification protocols vailable.Such that in literature,including certificate –based ,identity based identification[3][9]. According to that ,there are two phase,Proof and verify.In the stage of Proof ,a prover can prove their authorized identity to verifier.The verifier verify that prover identity based information and then proceed by either accept or reject.This protocol is mainly used in our system because user uploading file if user found some duplication on this particular file .This is most important protocol for secure user identity .

To provide data integrity ,the Azure stroage service stores the uploaded data MD5 checksum in the database and returns it to the user when user wants to retrieve the data.Amazon AWS computes the data MD5 checksum and e-mails it to the user for integrity checking .

3 SYSTEM MODEL

This section contains the three different entities: . 1. Cloud Management System 2. Virtual cluster administrations 3.user . Fig1. Depicts the architecture view of proposed deduplication system

1) Cloud Management System(CMS):This is entity that provides the storage service .This is main management system which control all Virtual Cluster Administration and user which is connected under the Virtual Cluster Administration.Here each user (s) tag only stored instead of store entire file . This is maintaining each users tag for whole file management system.

2) Virtual Cluster Administration (VCA):Its an entity which has the expertise and capabilities that client(s)/user(s) do not have and trusted to assess and expose the risk of cloud storage services on behalf the client request.

Here Webserver and Shibboleth [13] available.Shibboleth perform Single- Sign-On(SSO) for authorized user entry in this phase.Webserver will redirect unauthorized user (s) to shibboleth for verification.Tag was generated for each user based on attribute based .Through tag shibboleth allow the SSO.That generated tag was shared by CMS for verification purpose only.

3) User :Its also an entity who will access the massive data and controlled by central management system and Virtual Cluster Administration.User upload the file to VCA.According uploading the file ,tag generated which is stored in VCA as well as CMS for perfect secure file management system. Tag generation phase occurred here only.

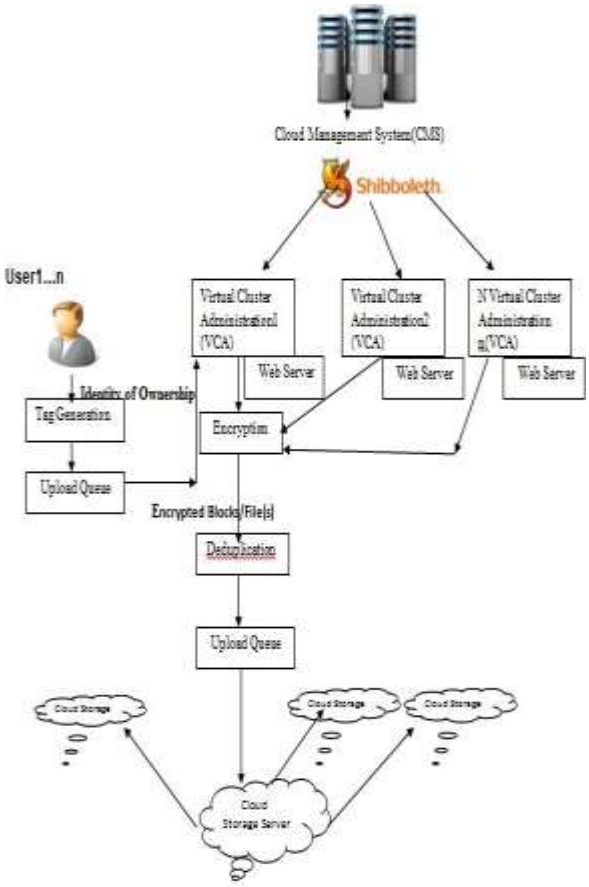

Fig1. Architecture view of proposed deduplication system.Optionally ,data on the back server can be replicated to multiple cloud storage in the back ground.

3.1 Adversary

Typically ,We assume that public and private cloud are both "honest but -curious" they will follow our proposed system.but users would try to access data either within or out off scopes of their privileges .

To illustrate this ,lets consider the following two scenarios. First ,assume that Alice , a company CFO ,stores the company financial data at a cloud storage service provided by Eve .And then Bob ,the company administration chairman ,downloads the data from the cloud .There are three important concerns in this procedure: Confidentiality,Integrity and Repudiation

Confidentiality : Eve is considered as an untrustworthy third party.Alice and Bob do not want reveal the data to Eve .

Integrity: As the administrator of the storage service ,Eve has capability to play with the data in hand.How can Bob be confident that the data he fetched from Eve are the same what was sent by Alice ? Are there any measures to guarantee that the data have not been tampered by Eve?

Repudiation: If Bob finds that the data have been tampered with,is there any evidence for him to demonstrate that it is Eve who should be responsible for the fault? Similarly ,Eve also needs certain evidence to prove her innocence .
Recently some reply from developer was ***"We wont lose your data ….we have a robust back up and recovery strategy but we are not responsible for you losing your own data -"***Obviously ,it is not persuasive to the potential customer to be confident with the service .
Confidentiality can not achieved by without adopting robust encryption schemes . But however ,the integrity and repudiation issues not handled well in the current cloud service platform . There are some linking missing between uploading and downloading sessions .
That leads some following questions:
Repudiation between users and VCA and Upload and Download sessions integrity

Repudiation between users and VCA: In case some data errors are occurred without transmission error means ,how can users and VCA prove their innocence and authorization?
Uploading and Downloading sessions integrity:
Since integrity in uploading and downloading phase are handled separately, how users and service providers download the same data content which is previously uploaded in system?
Is there mistake or error occurred means how to solve the sessions in uploading and downloading session?

4 SECURE DEDUPLICATION SYSTEM

There is a Authority Certificate (AC) issued by user and VCA and That AC copy is stored in CMS.The user and

VCA are using Key Sharing (KS),there are four solutions to bridge the mission link of data integrity between uploading and downloading procedures .

1) Neither AC nor KS
2) With KS without AC
3) With AC without KS
4) With both AC and KS

4.1 Neither AC nor KS:

Uploading file in our secure deduplication system :

1) User :sends the data to VCA with Tag generation and this known as Tag Generation by User(TGU)
2) VCA:verifies the data with Tag generation .If it is valid ,VCA sends back Tag Generation and this known as Tag Generation byVCA (TGVCA).

TGU is stored at the user side and TGVCA is stored at the VCA side .Then it is stored in cloud service provider without any problem.

Once uploading is finished ,both side agreed on the integrity of the uploaded data ,and each side owns TGU and TGVCA generated by opposite site.

Downloading file in our secure deduplication system:

1) User:Send request to VCA with Authorized identity proof (PoW)
2) VCA: Verifies the Authorized identity proof (PoW) .if it is valid ,the VCA send back the data with TGVCA to user .

User then verifies the data with TAG Generation. This things will update in CMS

4.2 With KS without AC

Uploading file in our secure deduplication sytem:

1) User: Sends the data to VCA with Tag Generation.
2) VCA:verifies the data with Tag Generation.If it is valid ,the VCA send back the Tag Generation.

VCA and user share Tag Generation with KS.

Then both sides agree on the integrity of the uploading data ,and they share the agreed Tag Generation ,which is used when disputation happens.

Downloading file in our secure deduplication system:

1) User :Send request to the VCA with Tag Generation
2) VCA:Verifies the request identity,if it is valid ,the VCA send back the data with Tag Generation
3) User verifies the data through Tag Generation.

When disputation happens the user or VCA can take the shared Tag Generation together recover it and prove his /her innocence.

4.3 With AC without KS

Uploading file in our secure deduplication system:

1) User : Sends the data to VCA along with Tag Generation by User (TGU).
2) VCA: Verifies the data with Tag Genearation,if it is valid ,the VCA send back the Tag Generation and TGVCA

TGU and TGVCA are send to AC

On finishing the uploading phase ,both sides agree on the integrity of the uploaded data , and AC owns their agreed Tag Generation.

Downloading file in our deduplication system:

1) User: Send request to VCA with Authorized identity proof (PoW).
2) VCA:verifies the request with PoW ,if it is valid ,the VCA send back the data with Tag Genration.

User verifies the data through the Tag Generation.

When disputation happens , the user or VCA can prove their innocence by presenting the TGU and TGVCA which are stored at the AC.

Similarly ,there are some special cases .when VCA is trustworthy ,only Tag Generation is needed.When user is trustworthy,only the TGVCA is needed;

4.4 With both AC and KS:

Uploading file in our secure deduplication system:

1) User : sends the data to VCA with Tag Generation.
2) VCA:Verifies the data with Tag Generation.

Both the user and VCA send Tag to AC.

AC verifies the two Tag .If they match ,the AC distributes Tag to the user and VCA by KS.

Both side agree on the integrity of the uploaded data and share the same Tag by KS and AC own their agreed Tag Generation.

Downloading file in our deduplication system:

1) User :Sends request to the VCA with PoW;
2) VCA:verifies the request identity ,if it is valid ,the VCA send back the data with Tag .

User verifies the data through Tag

Here are some special cases .when the VCA is trustworthy ,only the user needs Tag.When the user is trustworthy,only the VCA need Tag.

5 RELATED WORK

Yuan et al.[15] proposed a deduplication system in cloud storage to reduce the storage size of the tag for integrity check.Bellare [2]showed how to protect the data

confidentiality by transforming the predictable message into unpredictable message.Stanek et al[14] proposed some techniques that is for popular data and unpopular data .Li et al[12]addressed some key management by distributing keys across multiple servers after encryption the files .

Convergent Encryption:Xu et al [11] address the problem and showed a secure convergent encryption without considering key-management .It is known that some commercial cloud storage providers ,such as Bitcasa ,also deploy convergent encryption

Proof of Ownership(PoW): Halevi et al[10]proposed Proof of user authentication identity.Similarly Ng et al.[16]extended this same PoW but address the problem and how to management the key overhead. In our secure deuplication system based on attribute encryption .Therefore with out key overhead. And consider to be secured one by VCA and CMS.

6 EXPERIMENTAL RESULTS:

In this experimental result ,Fig 2 shows that how virtual machines node use the memory for storing the file .There are three categories available

1) Maximum
2) Average
3) And current

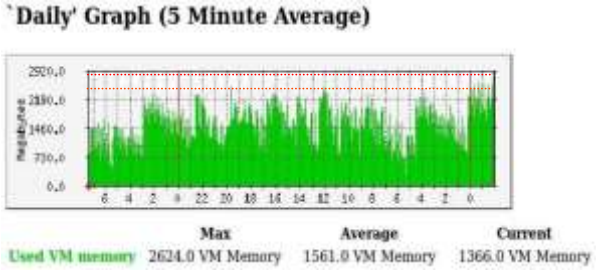

Figure 2. Timeline of the memory usage of the virtual machine.

Here there are some file are stored on daily basis.According maximum used Virtual Memory is 1460 MB to 2624.0 ,Average used Virtual memory is 730 MB to 1561 and current used virtual memory is 2190MB .Daily storing file and viewing file take memory usage is shown in experimental results.

7 FUTURE WORK

In our secure dedupliocation system,we consider three different categories such that CMS,VCA and user .Here group deduplication work and inter and intra group are considered.therefore generating Tag consume some amount of bandwidth .but not much more in cloud storage .In future reduce the generating tag space and bandwidth .Because tag usage is varied according transfer the file format such that .doc,.pdf,.jpeg,etc., Therefore we enhance fast transfer of any file from one group to another of user.

8 CONCLUSION

Our secure deduplication system is provided to protect the data from attack. Its make attacker job very complicated .because of generating Tag and PoW is making data more secure. In our system Cloud Management System (CMS) and Virtual Cloud Administration (VCA) having each user's file tag copy .Therefore transferring and storing file in cloud is very secure.Not visible to attack which means attack cannot find the path of target file.and CMS having CA for verifying owner identify by performing PoW.We showed that our secure deduplication system incurs minimal overhead compared to previous deduplication system.

9 ACKNOWLEDGMENTS

My Sincere Thanks to Mr.M.Newlin kumar sir and Mr. R.Nalla kumar sir ,mentors for providing their guidance and cooperation in this Research.

REFERENCES:

[1].J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer.Reclaiming space from duplicate files in a serverless distributed file system. In ICDCS, pages 617–624, 2002.

[2]. M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296–312, 2013.

[3]. M. Bellare, C. Namprempre, and G. Neven. Security proofs foridentity-based identification and signature schemes. J. Cryptology, 22(1):1–61, 2009.

[4]. R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman.Role-based access control models. IEEE Computer, 29:38–47, Feb 1996.

[5]. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis,and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.

[6]. D. Ferraiolo and R. Kuhn. Role-based access controls. In 15th NIST-NCSC National Computer Security Conf., 1992.

[7]. R. D. Pietro and A. Sorniotti. Boosting efficiency and security in proof of ownership for deduplication. In H. Y.

Youm and Y. Won, editors, ACM Symposium on Information, Computer and Communications Security, pages 81–82. ACM, 2012.

[8]. W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441–446. ACM, 2012.

[9]. M. Bellare and A. Palacio. Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks. In CRYPTO, pages 162–177, 2002.

[10]. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis ,and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.

[11]. J. Xu, E.-C. Chang, and J. Zhou. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In ASIACCS ,pages 195–206, 2013.

[12] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.

[13]. Shibboleth, http://shibboleth.internet2.edu, 2010.

[14]. J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure data deduplication scheme for cloud storage. In Technical Report, 2013

[15]] J. Yuan and S. Yu. Secure and constant cost public cloud Storage auditing with deduplication. IACR Cryptology ePrint Archive, 2013:149, 2013

[16]. W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441–446. ACM, 2012

International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 24 - 29, 2015, ISSN:- 2319–8656

User participation in ERP Implementation: A Case-based Study

Samwel Matende
Faculty of Computing and Information Management,
KCA University,
Nairobi, Kenya

Patrick Ogao
School of Information and Communication Technologies,
Technology University of Kenya
Nairobi, Kenya

Josephine Nabukenya
School of Computing and Informatics Technology,
Makerere University,
Kampala, Uganda

Abstract: Information Systems (IS), such as Enterprise Resource Planning (ERP) systems, are being developed and used in organizations to achieve their business goals and to enhance organizational effectiveness. The effect of user participation on successful systems development and implementation of ERP systems continues to be an area of interest to researchers. Common understanding has been that extensive user participation is not only important, but absolutely essential to system success. Even with this understanding of user participation as one of the critical factor in successful IS development and implementation, empirical studies have been unable to conclusively link user participation to systems success. This paper uses a private university as a case study to examine the role played by user participation in the implementation of an ERP system. In order to achieve its objective, this study adopted a mixed method where both qualitative and quantitative approaches were used in the collection of data. The results of the study reveal that user participation has a positive impact on the likelihood of ERP system success, user participation by choice is the best, user participation leads to better understanding of system requirements, the more participation the more the satisfied the users are, and participation builds support for the system during implementation. From our results we conclude that user participation in ERP system implementation is critical for successful implementation.

Keywords: *Enterprise Resource Planning, ERP systems, ERP implementation, User Participation.*

1. INTRODUCTION

Implementation of an ERP system is a complex IT-related social phenomenon with a large body of knowledge (Sarkera and Leeb, 2003). Amoako-Gyampah (2007) asserts that this implementation involves large expenditures, lengthy periods, and organizational commitment. An organization that decides to implement an ERP system is subjected to technological, information, business processes and people challenges. This implementation affects users at various levels of the organization since it cuts across all functional units. These users range from top management to low level users who use the system on their day-to-day operations. Earlier studies on ERP systems that focused on critical success factors, such as Al-Fawaz *et. al*. (2008), have identified user participation and involvement as one the important factors for successful ERP implementation.

The subject of user participation and involvement in implementation of information systems has been an area of interest to researchers and practitioners. Panorama consulting group (2010) claim that user involvement is one of the most cited critical success factor in ERP implementation projects. The result of involving users in the ERP implementation is a better fit between the resulting system and the business processes (Panorama Consulting Group, 2013). Users are invited to participate in an information system development (ISD) process because they have accumulated rich application domain knowledge through long period of exposure to their job context. User participation is advocated in order to discover users' needs and points of view, validate specifications, and hence build better IS for the organization (Esteves *et al.,* 2005). Other benefits include enhanced system quality, increased user knowledge about the system, greater user commitment and user acceptance (Harris and Weistroffer, 2008).

Ever increasing competition, expanding markets and enhanced customer expectations are among the challenges that organizations face today. To overcome these challenges, Enterprise Resource Planning (ERP) systems offer an integrated, enterprise-wide view of an organization's corporate information. According to Ibrahim *et al.* (2008, pp. 1), "an ERP software is a set of applications that links systems such as manufacturing, financial, human resources, data warehousing, sales force, document management, and after-sales service together, and helps organizations handle jobs such as order processing and production scheduling". This characteristics differentiates ERP systems from the traditional information systems that are considered to be information *silos* of various operational units of the organization. These *silos* are not integrated. Another distinction between ERP systems and the traditional information systems is the fact that majority of ERP systems are commercial of the shelf (COTs) systems which are bought and customized by the implementing organizations.

Due to their complexities, the implementation of ERP systems is a process of great complexity strongly involving the whole company and users at all levels of the organization. This implementation has many conditions and factors which potentially influence its success and the success of the entire project. One of these factors is user participation (Al-Fawaz *et. al*., 2008). ERP systems are complex pieces of software. Consequently, many such implementations have been difficult, lengthy and over budget, were terminated before completion, and failed to achieve their business objectives even a year after implementation (Somers and Nelson, 2004). The significance and risks of ERP make it essential that organizations focus on ways to improve ERP implementation.

Because of the promise of integration and facilitation on rapid decision-making, more organizations and institutions globally are implementing ERP systems (Markus and Tanis, 2000).

Along with this adoption, there has also been a greater appreciation of the challenges that arise from implementing these complex technologies. Al-Mashari (2003) asserts that many organizations are now adopting ERP systems making them today's most widespread IT solutions. For the benefits to be achieved, the ERP implementation should be successful. Many researchers have identified user participation as a critical success factor in the implementation of information systems and ERP systems. The way users participate during the development of a traditional information system is different from the way users participate in an ERP implementation. This paper investigates the role of user participation in ERP implementation using a private university as a case study.

The paper is divided into five remaining sections. Section 2 presents a review of related literature on user participation and past research on ERP implementation. Section 3 describes the methodology followed in collection and analysis of data. A brief description of the case study university and how an ERP system was implemented in this university is presented in section 4. The data on user participation in the ERP implementation process is presented in this section. Section 5 gives a brief discussion of the findings while section 6 concludes this paper.

2.REVIEW OF RELATED WORKS

Research on user participation in the past has focused on identifying the link between user participation or involvement and success in system development. No conclusive proof of this link has been offered by these past researches (Cavaye, 1995). Part of the reason for the inconclusive results can be traced back to the lack of clear definition of what user participation is. Prior to 1989, user participation and involvement were used interchangeably. Cavaye (1995) emphasizes that a clear definition of these terms will resolve ambiguities which might have brought about several interpretations of the research findings.

Traditionally, user participation was defined as user representatives' participation in IS development process (Ives and Olson, 1984), with reference to a series of specific activities undertaken by users (Baroudi *et al*., 1986). The term user participation in the information systems discipline used to be used interchangeably with user involvement till clear distinction was made between the two by Barki and Hartwick (1989; 1994).

Barki and Hartwick (1989) argue that user participation and user involvement are two distinct concepts with different meanings as used in other disciplines. It is worth noting that in the IS literature, these two concepts have been used interchangeably. In trying to differentiate these two concepts, Barki and Hartwick argue that *user participation* should be used to refer to those activities and behaviours of users and their representatives during the development process of an information system while on the other hand *user involvement* should be used to refer to the subjective psychological state that reflects the level of importance and personal relevance of the information system to users. We adopt this understanding of user participation in this study. We perceive user participation to be activities of users and their representatives during the development and implementation of an information system.

Despite the existence of various methodologies, there is a lack of measures to secure effective user participation. Qualitative research has found that users tended to play a marginal, passive, or symbolic role in IS development (for example, Beath and Orlikowski, 1994). For example, joint application development (JAD) fell short in promoting effective user participation, contrary to common expectation (Davidson, 1999; Gasson, 1999). Although participatory design (PD) places a strong emphasis on the cooperation between users and IS developers, there exists little evidence on the adaptability of this approach beyond the Scandinavian cultural settings (Carmel *et al*., 1993).

It is difficult for users to participate meaningful for the following reasons: (1) Led by IS staff, users tended to be drowned in technical materials (Davidson, 1999). (2) The language of communication between the two sides tends to be technical, often involving a great deal of jargons (Beath and Orlikowski, 1994; Davidson, 1999). And lastly, (3) Users are put into a passive position, lacking motivation for substantive participation (Wilson *et al*., 1997). These problems prevent users from participating in IS development in a meaningful and effective manner. Therefore, there is a clear need for further research on methods for effective user participation since users are the domain experts and carry a wealth of experience when it comes to their operation areas.

Few studies have been conducted on ERP implementation from the perspective of user participation. However, user participation issues were also touched upon in two other bodies of research on ERP implementation, albeit not as the main focus. The first one investigates critical success factors for ERP implementation. Secondly, user participation occasionally appears in recent research on ERP implementation team from the client side. Prior research has addressed the following themes: (1) the important role of absorbing ERP knowledge by the user team (Lorenzo *et al*., 2005; Robey *et al*., 2002); (2) the various types of ERP knowledge to be learned, such as the functionality of the software, idea of integration (Ko *et al*., 2005), and project management methods (Xu *et al.*, 2006); and (3) factors affecting the user team's absorption of ERP related knowledge (Ko *et al*., 2005).

Due to the complex nature of these systems, there have been reports of ERP implementation projects that do not succeed. Sumner (2000) states that there are a number of potential explanations for these failures. The failures can broadly be classified as human/organizational reasons such as lack of strong and committed leadership, technical reasons such as challenges or difficulties that arise from software customization and testing and economic reasons such as lack of economic planning and justification). Sumner (2000) further asserts that as much as each of these classes is important there appears to be a growing consensus among researchers that human factors are critical to the success of ERP projects.

In summary, the significance of user participation, as an important issue in ISD, has not been duly recognized in ERP implementation research. Research on ERP implementation from the perspective of user participation is lacking, and this could be an area of significant research contribution.

3. METHODOLOGY

A case study methodology was used in this study. The data collected for the purposes of this study comes from one case study conducted in a private university in Kenya. Qualitative data collection method (mainly semi-structured interviews) and quantitative data collection method (mainly a questionnaire were utilized. The interviewees were heads of departments, internal ERP project manager and IT

technicians. The questionnaire was administered using an online tool (SurveyMonkey) and it focused on the end users in various departments such as Finance, Human Resource, Registrar's office, Examinations and some teaching staff. The users who filled the questionnaire were identified by the IT Manager and the internal ERP project manager.

The survey questionnaire asked for information on user participation in the ERP implementations. ERP implementation has different phases and the users participated in either all or some of these phases. The questionnaire was four pages long and had a total of 16 questions addressing the various phases of ERP implementation. The questions in the survey required multiple responses while others were open-ended. The responses were encoded using a mix of check boxes (for multiple choices), radio buttons (for single choice) and open-ended answers. After the initial development of the survey questionnaire, it was sent to two ERP project leaders from our case study university and the ERP vendor. The primary objective was to test whether the instrument provided consistent and accurate information. Their responses were checked against the information collected during the case study. In addition, the questionnaire was checked by two ERP consultants. Based on the information provided by these experts, the instrument was fine-tuned and finalized.

The Case Study University

The case study was a private university in Kenya located in Nairobi which has adopted an ERP system and implemented several modules in several functional units. The university has broken down the ERP system into modules that handle several of its functional departments/units. These include finance, human resource, student management (admissions and examinations management). Other modules have been left out to be installed in the future. The university has four campuses which are all connected together and use the ERP system. At present, it employs in excess of 300 staff (both administration and teaching). This university is among the first private universities to implement an ERP system.

Before the implementation of the ERP system, the university had a legacy system that was not complex and was mainly used in the finance function. All other functional units used either a spreadsheet or a database application. All units were operating as information silos with no link or connection to other functional units. Prior to the implementation of the ERP system, the campuses were operating in isolation.

ERP Implementation in the Case Study Institution

The process of implementing the ERP system in the case study institution was triggered by the desire to have an integrated information system that will integrate the information generated from different functional units into one seamless system.

The university management set up an ERP Project committee which comprised of the ICT Director, an ERP Project Manager (a newly appointed staff) and representatives from the Academic and Administration divisions. This is the committee that was mandated by the university's management to spearhead the ERP adoption and implementation project. This committee commenced its work mid 2004.

The ERP lifecycle framework proposed by Esteves and Pastor (1999), figure 1 below, presents the main phases that the case study university followed in the implementation of the ERP system. A review of literature reveals that there is no consensus regarding the ERP lifecycle phases.

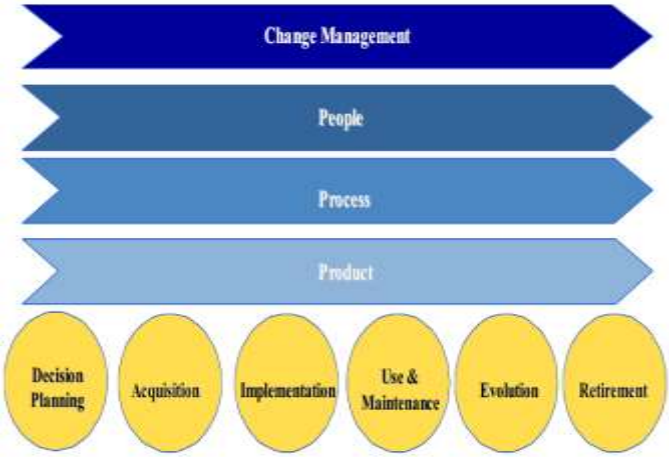

Figure 1: ERP Lifecycle Framework. *Source: Esteves and Pastor (1999)*

In the Adoption decision phase, the ERP Project committee reviewed the need for a new ERP system against the information needs of the university. This involved identifying the university strategic direction with regard to information systems solutions that best address the critical business challenges and improve the organizational strategy. The outcome of this phase is clear definition of system requirements, its goals and benefits, and an analysis of the impact of adoption at a business and organizational level. The Acquisition phase consists of the product selection that best fits the requirements of the organization. This aims at minimizing the need for customization during the implementation phase (Esteves and Pastor, 1999). Factors such as price, training and maintenance services are analyzed and, the contractual agreement played a major role in the identification of the ERP vendor. The committee settled on Microsoft Dynamics NAV as the ERP system to adopt. At this point, a team from the ERP vendor was constituted and liaised with the committee on the university side to continue with the ERP implementation.

During the Implementation phase, focus shifted from the ERP Project committee to the ICT department and the system developers from the vendor's side for the technical aspects of the implementation. This phase consists of the customization or parameterization and adaptation of the ERP package acquired according to the needs of the organization/institution (Esteves and Pastor, 1999). Other than the financial module, all the others required customization to meet the needs of the university. It is during this phase that the user participation was evident since they are considered to be the domain experts. This participation is discussed further in the next section.

The case university is currently in the Use and maintenance and the Evolution phases which involve the use of the product in a way that returns expected benefits and minimizes disruption and the integration of more capabilities into the ERP systems. The ERP system is under a maintenance contract which provides for any maintenance if and when it malfunctions and has to be corrected, special optimization

requests have to be met, and general systems improvements have to be made. The university has also added more modules onto the system such as the Procurement and is the process of adding Customer Relations Management (CRM).

4. FINDINGS

A total of fifteen users who had been identified by the IT manager and the internal ERP project manager as those that had participated in the various phases of the implementation process responded to the questionnaire. A web link was sent to them by email. As mentioned in the methodology section, these users come from different functional units of the university.

User participation in the ERP implementation in the case study university was evidence in all phases of the implementation. Users from different departments and sections of the university were selected to participate in the ERP implementation. These sections included finance, human resource, registrar office (which included Exams and Admission sections) and faculties. 29% participated in project definition, 62 % participated in requirements definition.

In terms of what capacity the users participated in the implementation 62% participated as end users who use the system in their daily operations while 38% participated as representatives of their departments/sections. A good example for this case was in the finance department where a team of users was identified by the department head to participate in the implementation process since the whole department would not be allowed to participate due to the critical nature of the department and provision of services to the clients.

With regard to the area that they participated in whether technical (which involved identification and purchase of the hardware and ERP software), social (human interface design, for example, design of forms, reports, etc) or module implementation (for example, participation in the implementation of a specific module such the finance or HR module), 36% of the users participated in the social aspect while 64% participated in module implementation. 30% of the users also participated in the testing of the system. The users who stated that they had participated in the social aspect were more interested with how the interfaces looked like and positioning of the menu items on the forms while the those that participated in module implementation were instrumental in the technical aspects of the system such as the modules meeting the requirements of the department.

In terms of communication from top management and ERP Project committee to the users is concerned, 53% stated that to a large extent users were adequately briefed about the ERP implementation process before it started, 82% were given an opportunity to perform various tasks relating to the implementation. 65% of users agreed that there was adequate communication between the ERP experts and users during the implementation process. With regard to reviewing of the work done by the ERP vendors, 56% of the users stated that they were given an opportunity to review the work done by the ERP vendors.

In assessing whether the ERP implementation was successful or a failure, 55% of the users stated that it was a success while 45% indicated that the implementation was average to below average. No user stated that the implementation was a total failure. 55% of the users are satisfied with the system most of the time while 42% are only satisfied some of the time when using the system. 15% of the users considered the implementation of the ERP system to be very successful, 40% consider the implementation to be successful, and 25% think it is average while 20% of the users believe that the implementation is below average.

5. DISCUSSION

In the case study university, there is evidence that there was effort by the university management to involve users in the ERP implementation process. The setting up of the ERP Project Committee to spearhead the implementation provided a form of project team that would engage users within the university and the experts from the ERP vendor team. The university has no policy on user participation when it comes to system development or implementation and hence not institutionalized.

An ERP system is a complex system that integrates the various functional units of an organization presenting uniform and real time information to these units. It comprises different modules that may be implemented at one go or in a phased approach. The ERP systems adopted by the university are a Commercial-Of-The-Shelf (COTS) system. The university adopted a phased approach where modules of some key functional units were implemented first. Users participated in different phases of the implementation with some participating in the project definition, others in requirements definition while others in the module implementation and testing.

Users were allowed to interact with the ERP experts from the vendor's team where their contributions were considered and taken seriously. This presented them with an opportunity to share their expertise and knowledge in their domain area. Departmental/section heads requested the users to participate voluntarily in the various meetings and sessions during the implementation.

The present study confirms the role played by user participation in ERP implementation. Users presented insights into their areas of operations which made it easier to identify the system requirements. During the meetings, the needs of the users were also discovered and incorporated into the implemented system. This was achieved through customization of some features of the system. Due to the participation, most users have accepted the system and are using it.

This study illustrates that user participation indeed contributes greatly to the success of ERP implementation. The successful elicitation of what were complex requirements led to a better understanding of both the business practices that the ERP system presents and the operations and duties performed by users.

6. CONCLUSION

The introduction of a new information system such as an ERP system will definitely change the way people work. These changes arise from new a platform, new and different interfaces, data entry is changed and report formats are different. Users often find these changes unnecessary and therefore refuse to accept them. One of the ways to address and reduce the impact of these changes is to encourage user participation in the implementation of ERP systems.

ERP implementations are expensive and complex undertakings, but once they are successfully implemented, significant improvements can be achieved such as easier

access to reliable information, elimination of redundant data and operations, reduction of cycle times, increased efficiency hence reducing costs (Zhang *et al*., 2003). This implementation differs from that of any traditional information system due to that it integrates different functional units of the organization. This leads to dramatic changes on how work is carried out, organizational structure and on the way people do their jobs. Most ERP systems are not built but adopted and thus they involve a mix of business process re-engineering (BPR) and package customization. This makes them unique and their implementation goes beyond technical concerns but also a socio-technical challenge since it affects how users perform their tasks.

ERP implementation differs from traditional systems development where the key focus has shifted from a heavy emphasis on technical analysis and programming towards business process design and human elements (Gibson *et al*., 1999). Unlike most home-grown legacy systems or those systems that are developed internally that were designed to fit individual working convention, ERP systems provide best practices, in other words generic processes and functions at their outset.

Aligning standard ERP processes with the organization's business process is considered to be an important step in the ERP implementation process (Botta-Genoulaz *et al*., 2005). Implementing a packaged ERP system inevitably changes the way people work. Successful implementation of an ERP system requires cooperation among different parties and departments.

In this paper, we investigated the contribution made by user participation in the ERP implementation process. A private university that implemented an ERP system was used as a case study. The implementation process followed the ERP lifecycle framework presented in figure 1 above. User participation positively impacted the implementation process and the majority of the users stated that the process was successful. Information systems are designed for use by users during their daily operations hence they are considered to be user-interfaced. This is also true of ERP systems which are designed to provide information processing capability to support the strategy, operations, management analysis, and decision-making functions in an organization. The user is at the center of an information system. Our study confirms that user participation is a critical to the success of the ERP implementation process.

There is a need, however, to investigate the role played by users in the process of customizing these commercially-of-the-shelf systems that have been designed for an educational institution. Educational institutions have different business processes unlike the manufacturing organizations. One of the limitations of this study was the fact that the adopted ERP system is not developed with university operations in mind. This system is strong on its financial modules. Another area for further research could be to investigate the role of user participation in internally developed ERP systems especially in Higher Education Institutions (HEIs).

In conclusion, we would like to reiterate the fact that ERP implementation is a complex IT-related social phenomenon. A substantial number of ERP implementations fail with a number of potential explanations for these failures presented. These failures, according to literature, may broadly be classified as human/organizational, technical, and economic. While each of these is important, there appears to be a growing consensus among researchers that human factors, more than technical or economic, are critical to the success of ERP projects.

7. REFERENCES

[1] Al-Fawaz, K., Al-Salti, Z. and Eldabi, T. (2008). *Critical Success Factors in ERP Implementation: A Review*. European and Mediterranean Conference on Information Systems 2008 (EMCIS2008) May 25-26, Al Bustan Rotana Hotel, Dubai

[2] Al-Mashari, M. (2003). Enterprise resource planning (ERP) systems: a research agenda. *Industrial Management & Data Systems,* Vol. 103/1, pp. 22-27

[3] Amoako-Gyampah, K. (2007). Perceived usefulness, User Involvement and behavioral intention: An empirical study of ERP implementation. *Computers in Human Behavior* Vol. 23, pp. 1232–1248

[4] Baroudi, J. J., Olson, M. H. and Ives, B. (1986). An Empirical Study of the Impact of User Involvement on System Usage and Information Satisfaction. *Communications of the ACM* (29:3), March 1986, pp. 232-238.

[5] Barki, H. & Hartwick, J. (1989). Rethinking the Concept of User Involvement. *MIS Quarterly*, March, pp. 53 – 63.

[6] Barki, H. and Hartwick, J., (1994). Measuring User Participation, User Involvement, and User Attitude. *MIS Quarterly*, 13:1, pp. 59 – 82.

[7] Beath, C., & Orlikowski, W. (1994). The contradictory structure of systems development methodologies: deconstructing the IS-User relationship in information engineering. *Information Systems Research*, 5(4), pp. 350-377.

[8] Botta-Genoulaz, V., Millet, P. (2006). An investigation into the use of ERP systems in the service sector. *International Journal of Production Economics*, 99 (1–2), 202–221.

[9] Carmel, E., Whitaker, R. D. and George, J. F. (1993). PD and Joint Application Design: a Trans-Atlantic comparison, *Communications of the ACM*, Vol. 36, (4), pp. 40-48.

[10] Cavaye, A. L. M. (1995). User Participation in System Development Revisited. *Information and Management.* Vol. 28, pp. 311 – 323.

[11] Davidson, E.J. (1999). Joint Application Design (JAD) in practice. *Journal of Systems and Software*, Vol. 45, pp. 215-223.

[12] Esteves J., Pastor J. (1999). *An ERP Lifecycle-based Research Agenda*. Proceedings of 1st International Workshop on Enterprise Management and Resource Planning: Methods, Tools and Architectures - EMRPS'99, pp 359-371.

[13] Esteves, J., Pastor, J. and Casanovas, J. (2005). Monitoring User Involvement and Participation in ERP Implementation Projects. *International Journal of Technology and Human Interaction*, 1 (14), p. 1 – 16.

[14] Gasson, S. (1999). The Reality of User-Centered Design. *Journal of End User Computing*, 11 (4), pp. 3 – 13

[15] Gibson, N.; Holland, C. and Light, B. (1999). A Case Study of a Fast Track SAP R/3 Implementation at Guilbert. *Electronic Markets*, June, pp.190-193.

[16] Harris, M. A. and Weistroffer, H. R. (2008). *Does User Participation Lead to System Success?* Proceedings of the Southern Association for Information Systems Conference, Richmond, VA,

USA.

[17] Ibrahim, A. M. S., Sharp, J. M., and Syntetos, A. A. (2008). *A Framework for the implementation of ERP to improve Business Performance: A Case Study.* European and Mediterranean Conference on Information Systems 2008 (EMCIS2008) May 25-26, Al Bustan Rotana Hotel, Dubai

[18] Ives, B. and Olson, M. H. (1984). User Involvement And MIS Success: A Review Of Research. *The Institute of Management Science*, Vol. 30, No. 5, pp. 586-603.

[19] Ko, D., Kirsch, L., King, W. (2005). Antecedents of Knowledge Transfer from Consultants to Clients in Enterprise System Implementations. *MIS Quarterly*, Vol 29 No. 1, pp. 59-85

[20] Lorenzo, O., Kawalek, P. and Wood-Harper, T. (2005). "Embedding the Enterprise System into the Enterprise: A Model of Corporate Diffusion", Communications of the Association for Information Systems (CAIS), (15), pp. 609-641.

[21] Markus, L. and Tanis, C. (2000). The Enterprise System Experience – from Adoption to Success. *Pinnaflex Educational Resources, Inc.*, Cincinnati, OH, pp. 173 – 207.

[22] Panorama Consulting Solutions 2010 ERP REPORT : *A Panorama Consulting Solutions Research Report.*

[23] Panorama Consulting Solutions 2013 ERP REPORT : *A Panorama Consulting Solutions Research Report.*

[24] Robey *et. al.*, (2002). Learning to implement enterprise systems: an exploratory study of the dialectics of change. *Journal of Management Information Systems*. v19 i1. 17-46.

[25] Sarkera, S. and Leeb, A. S. (2003). Using a case study to test the role of three key social enablers in ERP implementation. *Information & Management* Vol. 40, pp. 813–829.

[26] Somers, T. M. and Nelson, K. G. (2004). A taxonomy of players and activities across the ERP project life cycle. *Information and Management*, 41, pp. 257 – 278.

[27] Sumner, M. (2000). Risk factors in enterprise-wide/ERP projects. *Journal of Information Technology*, Vol. 15, pp. 317–327.

[28] Wilson, A., Bekker, M., Johnson, H. And Johnson, P. (1997). *Helping and hindering user involvement – A tale of everyday design*. Conference on human factors in computing systems (CHI) (Atlanta: ACM), pp. 221 -240.

[29] Xu, Q. and Ma, C., Zhang, C. and Su, M. (2006). "The mediation effect of transfer activities in ERP knowledge transfer". Proceeding of 15th International Conference on Management of Technology (IAMOT), Beijing, China, May 22 – 26.

[30] Zhang, L. *et. al.*, (2003). Critical Success Factors of Enterprise Resource Planning Systems Implementation Success in China. *Proceedings of the 36th Hawaii International Conference on System Sciences.*

International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 36 - 41, 2015, ISSN:- 2319–8656

Design of a Clinical Decision Support System Framework for the Diagnosis and Prediction of Hepatitis B

Adekunle Y.A
Department of Computer Science,
Babcock University,
Ilishan-Remo, Ogun State,
Nigeria

Abstract: This paper proposes an adaptive framework for a Knowledge Based Intelligent Clinical Decision Support System for the prediction of hepatitis B which is one of the most deadly viral infections that has a monumental effect on the health of people afflicted with it and has for long remained a perennial health problem affecting a significant number of people the world over. In the framework the patient information is fed into the system; the Knowledge base stores all the information to be used by the Clinical Decision Support System and the classification/prediction algorithm chosen after a thorough evaluation of relevant classification algorithms for this work is the C4.5 Decision Tree Algorithm with its percentage of correctly classified instances given as 61.0734%; it searches the Knowledge base recursively and matches the patient information with the pertinent rules that suit each case and thereafter gives the most precise prediction as to whether the patient is prone to hepatitis B or not. This approach to the prediction of hepatitis B provides a very potent solution to the problem of determining if a person has the likelihood of developing this dreaded illness or is almost not susceptible to the ailment.

Keywords: Hepatitis , Clinical Decision Support System (CDSS), Medical Decision Support System (MDSS), Artificial Intelligence (AI), K Nearest Neighbor (K-NN), Decision Trees (DT), Support Vector Machine (SVM) and Sequential Minimal Optimization (SMO)

1. INTRODUCTION

In recent times, the development of intelligent decision making applications is fast gaining ground. This concept is known as Artificial Intelligence (AI). Artificial Intelligence has different sub-fields which include expert systems, machine vision, machine learning and natural language processing amongst others.

A Decision Support System is an interactive computer-based system intended to help decision makers utilize data and models in order to identify and solve problems and make decisions [1]. According to the Clinical Decision Support (CDS) Roadmap project, CDS is "providing clinicians, patients, or individuals with knowledge and person-specific or population information, intelligently filtered or present at appropriate times, to foster better health processes, better individual patient care, and better population health."

A Clinical Decision Support System (CDSS) is an active knowledge system, where two or more items of patient data are used to generate case-specific recommendation(s) [2]. This implies that a CDSS is a decision support system (DSS) that uses knowledge management to achieve clinical advice for patient care based on some number of items of patient data. This helps to ease the job of healthcare practitioners, especially in areas where the number of patients is overwhelming.

Hepatitis B is a viral disease process caused by the hepatitis B virus (HBV). The virus is endemic throughout the world. It is shed in all body fluids by individuals with acute or chronic infection. When transmission occurs from mother to child or between small children during play, the infection nearly always becomes chronic. By contrast, when transmission occurs in adolescents/adults—usually via sexual contact, contaminated needles or other sharp objects, and less often from transfusion of blood products—the infection usually resolves unless the individual is immunocompromised.

Two billion people worldwide have serologic evidence of past or present HBV infection, and 350 million are chronically infected and at risk of developing HBV-related liver disease. Some 15–40% of chronically infected patients will develop cirrhosis, progressing to liver failure and/or HCC. HBV infection accounts for 500,000–1,200,000 deaths each year.

Health-care workers remain an at-risk group due to the risk of needlestick injury, and they should therefore all be vaccinated before employment.

Individuals chronically infected with HBV are at increased risk of developing cirrhosis, leading to hepatic decompensation and hepatocellular carcinoma (HCC). Although most patients with chronic HBV infection do not develop hepatic complications, there is the likelihood that serious illness can develop during their lifetime, and it is more likely to occur in men.

Every individual chronically infected with HBV provides an avenue for further cases to be prevented. It is expedient to take the time needed to educate patients and to explain the risks that the infection poses to the patients themselves and to others.

Hepatitis B vaccination is highly efficacious, and universal vaccination at a tender age is desirable. At the very least, vaccination should be offered to all individuals who are at risk. Pregnant women ought to be screened for hepatitis B before delivery, as this offers an opportunity to prevent another generation of chronically infected persons.

HBV-related liver injury is majorly caused by immune-mediated mechanisms, mediated via cytotoxic T-lymphocyte

lysis of infected hepatocytes. The precise pathogenic mechanisms responsible for the HBV-associated acute and chronic necroinflammatory liver disease and the viral and/or host determinants of disease severity have only recently been established [3]. The immune response of the host to HBV-related antigens is important in determining the result of acute HBV infection. The strength of the immune response of the host is critical for clearing the virus, but this simultaneously causes liver injury (i.e., a form of "hepatitis" manifested by a rise in transaminases occurs before clearance of the virus can be achieved). Those who become chronically infected are not able to sustain an immune response to HBV and thus undergo intermittent episodes of hepatocyte destruction (hepatitis).

Most studies of acute HBV infection are only initiated after the onset of symptoms, so that the critical early events following HBV infection go without notice. A recent study serially profiled the genomic changes during viral entry, spread, and clearance of the virus and showed that HBV does not induce any interferon-regulated genes in the early phase of the infection. Additionally, no genes were up-regulated or down-regulated in the lag phase of the infection or during the phase of viral spread. This suggests that HBV may not induce the intrahepatic innate immune response. Thus, HBV may be a "stealth" virus early in the infection.

When neonates are infected during childbirth if their mother is HBeAg-positive, immune tolerance is induced as the fetus becomes tolerized to the e antigen, a soluble viral protein that crosses the placenta in utero. This immune-tolerant phase goes on for years to decades. Children born to mothers who are HBeAg-negative but have ongoing viral replication more often develop an acute hepatitis in the neonatal period, which is cleared by the infant. However, the infectivity of many women who are HBeAg-negative is often very low, so that only about 20% transmit hepatitis B to their offspring.

Table 1. Acute hepatitis B infection: the risk of chronicity is related to age at primary infection (Source: Neth, 2006)

Outcome	Neonates	Children	Adults
Chronic infection	90%	30%	1%
Recovery	10%	70%	99%

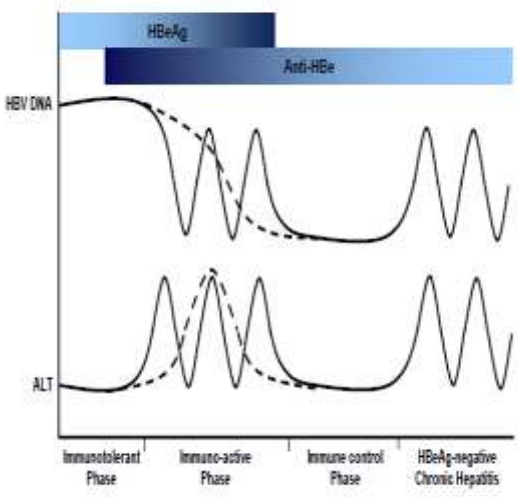

Figure 1. Chronic hepatitis B infection: phases of infection (Source: Buster, 2006)

Most cases of chronic hepatitis B in the reactivation phase are HBeAg-negative, but a few patients may be HBeAg-positive (Figure 2). The rates of progression to cirrhosis and hepatocellular carcinoma, with the associated mortality rates, are shown in Figure 2.

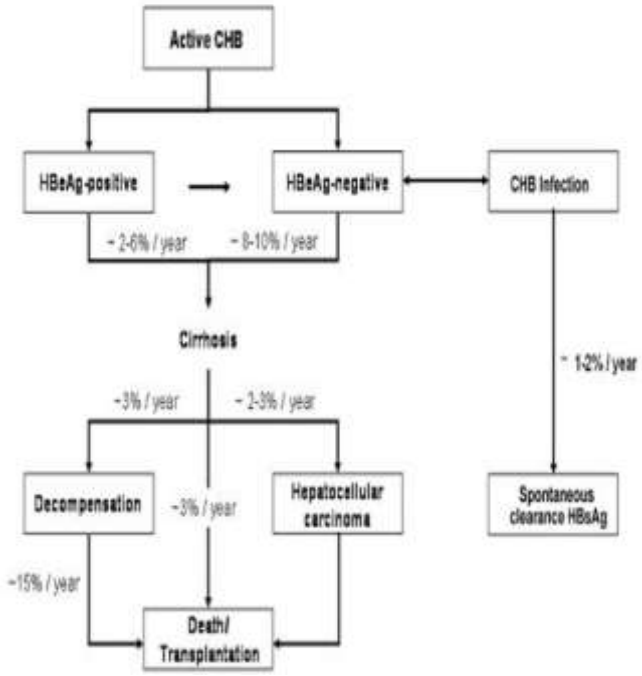

Figure 2. Progression to cirrhosis and hepatocellular carcinoma, with mortality rates (Source: Hepatol et al, 2003)

2. RELATED WORKS

2.1 Decision Support System for Heart Disease Based on sequential Minimal Optimization in Support Vector Machine

Here, Vadicherla & Sonawane (2013), the proponents of this system claim that computer based Medical Decision Support System (MDSS) can be useful for the physicians with its fast and accurate decision making process. They opined that predicting the existence of heart disease accurately, results in saving the lives of patients followed by proper treatment. Their objective was to present a MDSS for heart disease classification based on sequential minimal optimization (SMO) technique (which incorporated its features like high accuracy and high speed) in support vector machine (SVM). In using this method, they illustrated the UCI (University College Irvine) machine learning repository data of Cleveland heart disease database and consequently trained the SVM by using SMO technique. Hence, they also claim that given the ease of use and better scaling with the training set size, SMO is a strong candidate for becoming the standard SVM training algorithm. Training a SVM requires the solution of a very large QP (Quantum Platform) optimization problem. SMO algorithm breaks this large optimization problem into small sub-problems. Both the training and testing phases give the accuracy on each record. The results proved that the MDSS is able to carry out heart disease diagnosis accurately in a fast way and it was reported to show good ability of prediction on a large dataset.

2.2 Data Mining in Clinical Decision Support Systems for Diagnosis and Treatment of Heart Disease

According to Amin, Agarwal & Beg (2013) medical errors are both costly and harmful. Medical errors cause thousands of deaths worldwide each year. Hence, a clinical decision support system (CDSS) would offer opportunities to reduce medical errors as well as to improve patient safety. They affirm that one of the most important applications of such systems is in diagnosis and treatment of heart diseases (HD). This is because statistics have shown that heart disease is one of the leading causes of deaths all over the world (CDC Report). Data mining techniques have been very effective in designing clinical support systems because of its ability to discover hidden patterns and relationships in medical data. Here, the proponents also undertook a comparative analysis of the performance and working of six CDSS systems which use different data mining techniques for heart disease diagnosis. They conclude by asserting based on their findings that there is no system to identify treatment options for Heart disease patients. They further claimed that in spite of having a large amount of medical data, it lacked in the quality and the completeness of data thereby creating the need for highly sophisticated data mining techniques to build up an efficient decision support system. They claim that even after doing this, the overall reliability and generalization capability might still be questionable. Hence, the need to build systems which will be accurate, reliable as well as reduce cost of treatment and increase patient care. More so, the building of systems which are understandable and which could enhance human decisions are very germane.

2.3 An Intelligent Decision Support System for the Operating Theater

In 2013 Sperandio, Gomes, Borges, Brito and Almada-Lobo asserted that decision processes inherent in operating theatre organization are often subjected to experimentation, which sometimes lead to far from optimal results. They further affirm that the waiting lists for surgery had always been a societal problem, with governments seeking redress with different management and operational stimulus plans partly due to the fact that the current hospital information systems available in Portuguese public hospitals, lack a decision support system component that could help achieve better planning solutions. As such they developed an intelligent decision support system that allows the centralization and standardization of planning processes which improves the efficiency of the operating theater and tackles the fragile situation of waiting lists for surgery. The intelligence of the system is derived from data mining and optimization techniques, which enhance surgery duration predictions and operating rooms surgery schedules.

2.4 Decision Support System for the Diagnosis of Schizophrenia Spectrum Disorders.

In 2013, Kahn developed a decision support system for the diagnosis of schizophrenia spectrum disorders.

The development of this system is described in four-stages: knowledge acquisition, knowledge organization, the development of a computer-assisted model, and the evaluation of the system's performance. The knowledge is extracted from an expert through open interviews. These interviews aimed at exploring the expert's diagnostic decision making process for the diagnosis of schizophrenia. A graph methodology was employed to identify the elements involved in the reasoning process. Knowledge was first organized and modeled by means of algorithms and then transferred to a computational model created by the covering approach. The performance assessment involved the comparison of the diagnosis of 38 clinical vignettes between an expert and the decision support system. The results showed a relatively low rate of misclassification (18-34%) and a good performance by the decision support system in the diagnosis of schizophrenia, with an accuracy of 66-82%.

3. CLINICAL DECISION SUPPORT SYSTEM (CDSS)

The clinical decision support system is another example of a knowledge based system. A clinical decision support system is an active knowledge system where two or more items of patient data are used to generate case specific recommendations [2].

3.1 Target Area of care

CDSSs assist doctors in assessing various clinical issues from accurate diagnosis of a particular disease to the treatment of the disease. The general target areas of care for CDSS are:

Preventive care which has to do with screening and disease management

Diagnosis which is done based on the patients' signs and symptoms

Follow-up management which has to do with frequent checkups

Hospital Provider Efficiency [8].

3.2 System Design

The system design for CDSS will usually include the following subsystems:
Communication which handles notification and alerts
Knowledge discovery which deals with rules and regulations
Knowledge repository which contains problem solving knowledge [9].

3.3 Factors leading to successful CDSS implementation

The following under listed factors lead to the successful implementation of CDSS:
Simple, user friendly interface
Automated decision support
Timely result
Workflow integration
Continuous Knowledge-base and update support [10].

4. PATTERN CLASSIFICATION METHODS

Pattern classification refers to the theory and algorithms of assigning abstract objects into distinct categories, where these categories are typically known in advance. For this research, the pattern classification methods considered are Decision Trees (DTs), K-Nearest Neighbor (KNN), Naïve bayes Classifier and Support Vector Machine (SVM).

4.1 Decision Trees

A decision tree consists of a root node, branch nodes and leaf nodes. The tree begins with a root node, then further splits into branch nodes and each node represents a choice among various alternatives. The tree then terminates with leaf nodes which are un-split nodes that represent a decision [11]. The classification of decision trees are carried out in two phases:
Tree Building or top down: This is computationally intensive and requires the tree to be recursively partitioned until all the data items belong to the same class.
Tree pruning or bottom top: It is conducted to improve the prediction and classification of the algorithm and minimize the effects of over-fitting which may lead to misclassification of errors [12].
Some notable decision tree algorithms include Classification and Regression Trees (CART), Iterative Dichotomiser 3 (ID3), C4.5 and C5.0.
The advantages of decision trees include:

- They are easy to interpret and comprehend
- They can handle both metric and non-metric data as well as missing values which are frequently encountered in clinical studies.
- Little data preparation is required since data does not need to be normalized.
- They can handle data in a short time frame.
- They can be developed using common statistical techniques.

The disadvantages associated with decision trees include:

- They can over fit the data and create complex trees that may not generalize well.
- A small change in the size of a dataset could result in a completely different tree

4.2 K-Nearest Neighbor

K-Nearest Neighbor (k-NN) is instance based learning for classifying objects based on closest training examples in the feature space. It is a type of lazy learning where the function is only approximated locally and all computations are deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. If k=1, then the object is simply assigned to the class of its nearest neighbor. The k-NN algorithm uses all labeled training instances as a model of the target function. During the classification phase, k-NN uses a similarity-based search strategy to determine a locally optimal hypothesis function. Test instances are compared to the stored instances and are assigned the same class label as the k most similar stored instances.

4.3 Bayes Classifier

A Bayesian network is a model that encodes probabilistic relationships among variables of interest. This technique is generally used for intrusion detection in combination with statistical schemes, a procedure that yields several advantages, including the capability of encoding interdependencies between variables and of predicting events, as well as the ability to incorporate both prior knowledge and data. However, a serious disadvantage of using Bayesian networks is that their results are similar to those derived from threshold-based systems, while considerably higher computational effort is required.

4.4 Support Vector Machine

Support Vector Machines have been proposed as a novel technique for intrusion detection. An SVM maps input (real-valued) feature vectors into a higher-dimensional feature space through some nonlinear mapping. SVMs are developed on the principle of structural risk minimization. Structural risk minimization seeks to find a hypothesis (h) for which one can find lowest probability of error whereas the traditional learning techniques for pattern recognition are based on the minimization of the empirical risk, which attempt to optimize the performance of the learning set. Computing the hyper plane to separate the data points i.e. training an SVM leads to a quadratic optimization problem. The implementation of SVM intrusion detection system has two phases which are training and testing. SVMs can learn a larger set of patterns and be able to scale better, because the classification complexity does not depend on the dimensionality of the feature space. SVMs also have the ability to update the training patterns dynamically whenever there is a new pattern during classification.

Table 1. Comparison of Some Pattern Classification Algorithms (Source: Patel et al, 2012)

Classifier	Method	Parameters	Advantages	Disadvantages
Support Vector Machine	A support vector machine constructs a hyper plane or set of hyper planes in a high or infinite dimensional	The effectiveness of SVM lies in the selection of kernel and soft margin parameters. For kernels, different pairs of	1. Highly Accurate 2. Able to model complex nonlinear decision boundaries 3. Less prone to over fitting than other methods	1. High algorithmic complexity and extensive memory requirements of the required quadratic programming in large-scale

	space, which can be used for classification, regression or other tasks.	(C, γ) values are tried and the one with the best cross-validation accuracy is picked. Trying exponentially growing sequences of C is a practical method to identify good parameters.		tasks. 2. The choice of the kernel is difficult 3. The speed both in training and testing is slow.
K Nearest Neighbour	An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its *k* nearest neighbours (*k* is a positive integer). If *k* = 1, then the object is simply assigned to the class of its nearest neighbour.	Two parameters are considered to optimize the performance of the kNN, the number k of nearest neighbour and the feature space transformation.	1. Analytically tractable. 2. Simple in implementation 3. Uses local information, which can yield highly adaptive behaviour 4. Lends itself very easily to parallel implementations	1. Large storage requirements. 2. Highly susceptible to the curse of dimensionality. 3. Slow in classifying test tuples.
Bayesian Method	Based on the rule, using the joint probabilities of sample observations and classes, the	In Bayes, all model parameters (*i.e.*, class priors and feature probability distributions) can be	1. Naïve Bayesian classifier simplifies the computations. 2. Exhibit high accuracy and speed	1 The assumptions made in class conditional independence. 2. Lack of available probabilit
	algorithm attempts to estimate the conditional probabilities of classes given an observation.	approximated with relative frequencies from the training set.	when applied to large databases.	y data.
Decision Tree	Decision tree builds a binary classification tree. Each node corresponds to a binary predicate on one attribute; one branch corresponds to the positive instances of the predicate and the other to the negative instances.	Decision Tree Induction uses parameters like a set of candidate attributes and an attribute selection method.	1. Construction does not require any domain knowledge. 2. Can handle high dimensional data. 3. Representation is easy to understand. 4. Able to process both numerical and categorical data.	1. Output attribute must be categorical. 2. Limited to one output attribute. 3. Decision tree algorithms are unstable. 4. Trees created from numeric datasets can be complex.

5. METHODOLOGY

A very comprehensive dataset (Velvet, 2008) consisting of 100,000 instances compiled from the UCI (University of California, Irvine) data repository was used. This dataset was translated into the Attribute Relational File Format (ARFF) which is one of the file formats recognized by the WEKA (Waikato Environment for Knowledge Analysis) software in which the distinct genotypic attributes used for this work were highlighted.

The dataset was then induced with Classification algorithms namely C4.5 decision trees, Support Vector Machine (SVM), K-Nearest neighbor algorithm and Bayes Classifier Algorithm. The Classification algorithms were evaluated using the Waikato Environment for Knowledge Analysis software version 3.6.7 based on the percentage of correctly classified instances with the C4.5 decision trees having 61.0734%, the Support Vector Machine (SVM) algorithm had 50.0515%, the Bayes Classifier Algorithm had 50.2045% and the K-Nearest Neighbor algorithm had 50.1235%. Sequel to the result obtained from this evaluation, the C4.5 decision trees turn out as the Classification algorithm with the highest accuracy for this research. Thereafter, a decision tree program was written in Java with 38 lines of code for the core program

to implement the C4.5 decision tree algorithm that will provide the requisite intelligence for this Clinical decision support system and help it make the right decisions promptly when supplied with patient information. The C4.5 decision tree algorithm will be embedded in the classification/prediction algorithm section of the clinical decision support system.

6. CONCLUSION AND RECOMMENDATION

This research work finds its significance in all parts of the world where people live with the health challenge caused by the hepatitis B virus, thus it is very germane as it provides a sort of panacea to the eventual development of the condition known as hepatitis B for people who are susceptible to the condition, hence they can be aware of their susceptibility ahead of time and can be able to take the necessary precautionary measures to forestall their development of the illness, thus saving them from the trauma they would have inevitably suffered.

The research is a milestone in the sub-field of health informatics as it provides a readily available Clinical Decision Support System to serve as a reliable assistant to the medical practitioners that are more often than not burdened by the overwhelming and seemingly intimidating number of patients they need to attend to routinely. This has culminated in a lot of fatal errors on the part of the medical practitioners which has led to the loss of innocent lives hence, the introduction, consequent adoption and deployment of this Knowledge Based Intelligent Clinical Decision Support System for the prediction of hepatitis B becomes expedient especially in the third world countries, the vast majority of who lag behind in terms of technological innovations and advancement and as a result are alien to the terrific results gotten from the use of these clinical decision support systems.

For further work another enthusiastic researcher can go a step further in this work by introducing other highly efficacious algorithms that can be used alongside the C4.5 decision tree algorithm used in this work, so as to have a hybrid system that will take decisions faster and generate more accurate decisions than those that will be given by the proposed system.

7. REFERENCES

[1] Power, D.J. (1999). Decision Support Systems Glossary.http://DSSResources.COM/glossary

[2]Chen, J.Q & Lee, S.M. (2002). An exploratory cognitive DSS for strategy decision making. *Elsevier Science B.V.*

[3] Naumann, H, Scott R.M, Snitbhan R, Bancroft W.H, Alter H.J & Tingpalapong, M (2006). Experimental transmission of hepatitis B virus by semen and saliva. *International Journal of Infectious Diseases*, 23(8), 27-35.

[4] Vadicherla, D. & Sonawane, S. (2013). Decision support system for heart disease based on sequential minimal optimization in support vector machine. *International Journal of Engineering Sciences & Emerging Technologies*, 4(2), 19-26.

[5] Amin, S.U, Agarwal, K & Beg R. (2013). Data mining in clinical decision support system for diagnosis, prediction and treatment of heart disease. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 2(1), 56-67.

[6] Sperandio F, Gomes C, Borges J, Brito A.C &Almada-Lobo B. (2013). An intelligent Decision Support System for the Operating Theatre: A Case Study. *Automation Science and Engineering, IEEE Transactions on Robotics & Control Systems*, 99.

[7] Kahn, R., Perkins, D., Lieberman, J.(2012). Predictors of treatment response in patients with first-episode prostate cancer disorder. *The British Journal of Gynecology*, 185(1),18-24.

[8] Berner, E.S. (2009). Clinical Decision Support Systems: State of Art. Rockville: AHRQ Publication 12(5), 90-134.

[9] Frize, M. (2005). Conceptual Framework of Knowledge Management for Ethical Decision making Support in Neonatal Intensive Care. *IEEE Transactions of Information Technology in Biomedicine*, 9(7), 205-215.

[10] Peleg, M., & Tu, S. (2011). Decision Support, Knowledge Representation and Management in Medicine. Stanford Centre for Biomedical Informatics Research. Last accessed: December 17, 2011 from http://bmir.stanford.edu/file_asset/index.php/1009/SMI-2006-1088.pdf

[11] Peng, W., Chen J. & Zhou H. (2006). An Implementation of ID3-Decision Tree Learning Algorithm. University of New South Wales. Last accessed: December17, 2011 from http://web.arch.usyd.edu.au/~wpeng/DecisionTree2.pdf.

[12] Anyanwu, M.N., & Shiva, A.G., (2009). Comparative Analysis of serial decision tree classification algorithms. *International Journal of Computer Science and Security*, 3(3), 230-240.

[13] Neth, Q, & Alter M (2006). Epidemiology of hepatitis B in Europe and worldwide. *International Journal of Hepatology*, 39(4), 66-69.

[14] Buster, R, & Hyams K.C (2006). Risks of chronicity following acute hepatitis B virus infection. *International Journal of Clinical Infectious Diseases*, 20(3), 992–1000.

[15] Hepatol, O.M, Orito, E, Mizokami, M, Sakugawa, H, Michitaka, K, Ishikawa, K & Ichida, T (2003). A case-control study for clinical and molecular biological differences between hepatitis B viruses of genotype B and C. *Canadian Journal of Hepatology*, 14(7), 908-921

International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 42 - 47, 2015, ISSN:- 2319–8656

A Hybrid Prediction System for American NFL Results

Anyama Oscar Uzoma
Department of Computer Science
Faculty of Physical and Applied Sciences
University of Port Harcourt
Rivers State, Nigeria

Nwachukwu E. O.
Department of Computer Science
Faculty of Physical & Applied Sciences
University of Port Harcourt
Rivers State, Nigeria

Abstract: This research work investigates the use of machine learning algorithms (Linear Regression and K-Nearest Neighbour) for NFL games result prediction. Data mining techniques were employed on carefully created features with datasets from NFL games statistics using RapidMiner and Java programming language in the backend. High attribute weights of features were obtained from the Linear Regression Model (LR) which provides a basis for the K-Nearest Neighbour Model (KNN). The result is a hybridized model which shows that using relevant features will provide good prediction accuracy. Unique features used are: Bookmakers betting spread and players' performance metrics. The prediction accuracy of 80.65% obtained shows that the experiment is substantially better than many existing systems with accuracies of 59.4%, 60.7%, 65.05% and 67.08%. This can therefore be a reference point for future research in this area especially on employing machine learning in predictions.

Keywords: Data Mining, Hybrid System, K-Nearest Neighbour, Linear Regression, Machine Learning, National Football League

1. INTRODUCTION

Predicting the outcome of events is of interest to many, ranging from meteorologists, to statisticians, the media, to financial experts, to economists, to clubs, to merchants, to fans, to pundits and to betting markets (bookies). In the past, involvement in prediction was a leisure activity with elements of luck and experience inter operating. Now, predicting events outcome has become interesting as a research problem, in part due to its difficulty; this is because prediction outcome is dependent on many intangible or human induced factors. Predicting the outcome of event has also become a mega business but even armed with all these expertise in analyzing past data, it is very hard to predict the exact outcome of range of events.

Looking at games prediction as a sub-domain in predictions, the world has yielded huge profits and investments from these possible games outcomes. For example, NFL (National Football League) football is arguably the most popular games in North America. Over the past two decades alone, NFL has truly become America's game, with millions of people watching NFL games live on television at home and fans going to the stadium. With emerging facts that betting market in the United States accumulates nearly $1B per year on football games [1], hence, it implies that investments in forecasting outcomes in this area will be a worthy venture.

Countless number of people, computing tools and even internet websites claim they know or they can predict the possible win, lose and draw outcomes of future games. Their prediction results also come in varieties and some degree of bias, from the most accepted to the least accepted. With an avalanche of different opinions about prediction out there, the question becomes, which of these are actually predicting correctly?

With this research work, a completely data driven objective system was designed to predict the outcome of future games, purely for academic and business purposes.

2. RELATED WORK

A large number of literatures have been dedicated to the development of goal modeling, result modeling, ratings and rankings for games prediction. These works include:

[2], developed a Logistic Regression/Markov Chain Model for NCAA Basketball, in their work the National College Championships was used as their case study. Markov chain model was used for teams' ranking, the underlying model implements a chain with one state for each team. The intuition is that state transitions are like the behavior of a hypothetical voter in one of the two major polls. The current state of the voter corresponds to the team that the voter now believes to be the best. At each time step, the voter evaluates his judgment in the following way, given that he currently believes team i to be the best, he picks (at random) a game played by team i against some opponent j. With probability p, the voter moves to the state corresponding to the game's winner; with probability (1 – p), the voter moves to the losing team's state. The Logistic regression model used home-and-home conference data to estimate an answer questions from the existing problem. Good prediction accuracy was obtained with limitations on the poor ranking for losing teams and approach used only basic data.

[3], worked on A Quantitative Stock Prediction System based on Financial News. In their work the discrete stock price prediction using a synthesis of linguistic, financial and statistical techniques to create the Arizona Financial Text System (AZFinText) was done. The major objective of the project was to provide predictions for stock market using statistical data gathered from financial news. The lines of research approach used were Mean Squared Error (MSE), visualization tools and Machine Learning Techniques. Prediction accuracy of 71.2% was obtained with a Simulated Trading return of 8.50%.

[4], proposed a modified Least Squares approach incorporating home field advantage and removing the influence of margin of victory on ratings, identified key attributes of any ranking system and modeled these novel features into the new system. A prediction accuracy of 70.1 percent was obtained and the limitation of the approach is the fact that the modeled data was linear in nature.

[5], used simple regression-based technique to predict the outcome of football matches. The model investigated the linear relationship that exists between the variables and data sets. Number of games played, Scoring margin (average points scored per game minus average points, yielded per game, despite the BCS decree that margin of victory not be used for computer ratings, just to gauge the importance of scoring margin as a predictor), Offensive yardage accumulated per game, Offensive first downs per game, Defensive yardage yielded per game, Defensive first downs yielded per game, Defensive touchdowns yielded per game, Turnover margin (takeaways minus giveaways), Strength of schedule. Limitation of this system is that it is linear in nature, poor prediction accuracy of 59.4 percent.

Table 1: Summary of Related Works

AUTHOR(S)	TITLE OF RESEARCH WORK	FEATURES/ TEHNIQUES	EXPERMENT DATA SETS USED	PRED ACC	OBSERVED ADVANTAGES	OBSERVED PITFALLS
[6]	Ocean Model, Analysis and Prediction System	Root Mean Square Error	Daily height anomaly data	Good	Real time observation system	Complex approach
[7]	Advanced Regional Prediction System (ARPS)	Non-hydrostatic moel techniques Wth Perl Prog Lang	Snow assessment parameters, cloud access	Good	Real time parameters	Complex model
[8]	Prediction Model Study of Basketball Simulation Competition Result Based on Homogeneous Markov in NCAA	Markov Chain	Pre-season data	Good	Good prediction accuracy Simulation of pre-season results	Few dataset was obtained
[3]	A Quantitative Stock Prediction System based on Financial News	Mean Squared Error (MSE), Visualization tools and Machine Learning Techniques	Stock market using statistical data gathered from financial news	71.2 Percent	Simulated Trading return of 8.50% Good prediction accuracy	Complex model

3. ANALYSIS OF EXISTING SYSTEM

The existing system is the research work done by Brady T. and Madhur L. 2008. Their work involved the use of a straightforward application of linear modeling in the development of a predictive model for the outcomes of college football bowl games. The implemented techniques identifies important team-level predictors of actual bowl outcomes in 2007-2008 using real Football Bowl Subdivision (FBS) data from the completed 2004-2006 college football seasons. Given that Bowl Championship Series (BCS) ratings was used to determine the teams most eligible to play for a national championship and a playoff system for determining a national champion.

Their approach uses Linear Regression based technique to predict the outcome of football matches. The model investigated the linear relationship that exists between the variables and data sets.

3.1 Linear Regression

The existing system develops a model using a linear Multiple Regression approach which implies that more than one predictor variable is available and the linear components represents the regression coefficients being additive. The algorithm below represents the multi linear regression approach which provides a rating output.

Algorithm

Input

Attributes X1, X2…. Xn

Main process algorithm

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \varepsilon$

$\beta 0$ = intercept

$\beta 1 \beta \rho$ = regression coefficients

ε = residual standard deviation

Output

Y= Dependent variable (Predicted Result used for rating)

Features used

The following features were used: Scoring margin (average points scored per game minus average points yielded per game, despite the BCS decree that margin of victory not be used for computer ratings, just to gauge the importance of scoring margin as a predictor, Offensive yardage accumulated per game, Offensive first downs per game, Defensive yardage yielded per game, Defensive first downs yielded per game, Defensive touchdowns yielded per game, Turnover margin (take-aways minus give-aways), Strength of schedule (as computed by Jeff Sagarin for USA Today).

Limitations of the existing system

The following limitations were observed with the existing system. Over fitting of data, Poor prediction accuracy, only linear in nature, model cannot be trained and does not provide generalization to the prediction problem.

Prediction Accuracy: A prediction accuracy of 59.4 percent was obtained.

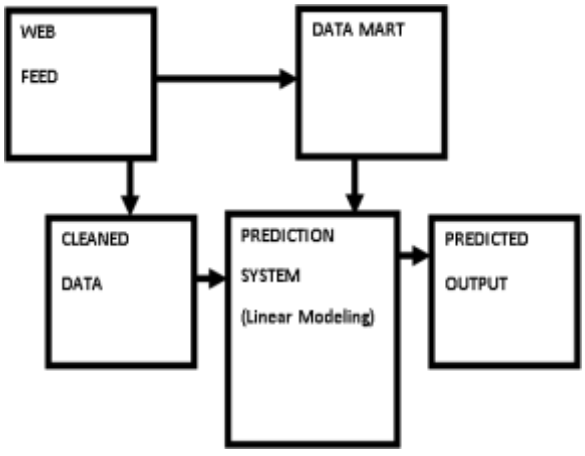

Figure 1. Brady T. and Madhur L., (2008)

(Existing System)

4. ANALYSIS OF PROPOSED SYSTEM

In the proposed system, machine learning algorithms were developed to out-perform the existing system. The proposed model framework is a hybrid of **Linear Regression Technique and K-Nearest Neighbour Technique**, which employs an objective supervised learning method. A major consideration in the choice of a hybrid system is that winning occurs in a variety of ways which in turn affects the statistics of these games. A quick glance at the various games statistics does not correctly provide the winner of a game, although it does lend some insight. In fact, there are countless examples of games in which statistics favor a team to win a game but the team eventually lost that game. Hence this indicates that there is no linear mapping method in which a winner can be chosen based solely on a group of statistics.

The hybrid system could be used to perform non-linear mapping based on a variety of relevant statistics, hence for this project the dataset to be considered will be available from the games portal. The importance and weight of each statistic must be determined prior to making a prediction using linear regression as this will provide appropriate statistical weights.

The K-Nearest Neighbour will provide classification of already weighed features. Results from trained prediction set will be applied to unseen games. The resulting hybrid model provides an optimized model which in turn will yield good results with good prediction accuracy with big implication on the various dependents of the results.

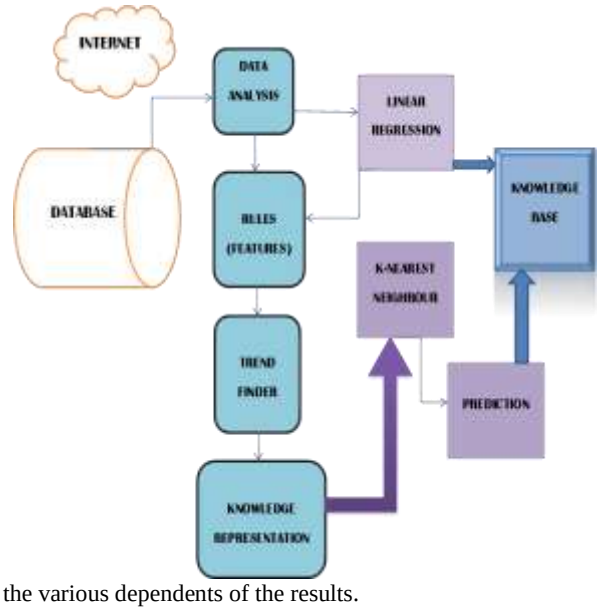

Figure 2. Proposed Hybrid Model

In designing the hybrid system, the following steps will be employed:

Step 1: Problem definition

The hybrid system will provide a correct understanding of the existing problem. Here the understanding will be broken into the project objectives and the requirements.

Step 2: Data collection and pre-processing

For the purpose of this hybrid system, there is need to acquire data from NFL box score. The dataset location on the server will be downloaded automatically from pro-football using Google's URL crawling tool and Microsoft Excel Web Extraction Feature then preprocessed with excel to acquire the right features. The dataset used is NFL games statistics for week one week 16 in the 2013 season.

Step 3: Modeling

This phase is the core of the hybrid system and will be divided into two sub steps:

i) Build model

ii) Execute model

Build

In this phase, the linear regression technique and K-Nearest Neighbour techniques will be developed using the features sets.

Execute

The resulting value of range -1 to +1 from the Linear Regression Technique will provide attributes that affects or contributes to the prediction results. The results will be used as a basis for the K-Nearest Neighbour model.

Figure 3. Algorithm of Implementation (Proposed Hybrid Model)

4.1 Features used

The following set of features will be used for the hybrid proposed system.

PtsW: Points Scored by the winning team

PtsL: Points Scored by the losing team

W#: Week number in season (Road)

YdsW: Yards Gained by the winning team

YdsL: Yards Gained by the road team

TOW: Turnovers by the winning team

TOL: Turnovers by the losing team

Tm: Points scored

Tm: Points scored

Rec: Team's record following this game (Streak)

WLD: Win, loss, draws percentage of the home team

WLD: Win, loss, draws percentage of the road team

TotYd: Total Yards Gained on Offense

TotYdL: Total Yards gained on defense

PassY : Total Yards Gained by Passing (includes lost sack yardage)

RushY: Total Rushing Yards Allowed by Defense

Sp. Tms: Special teams

Offense: Offense

Defense: Defense

Rating: Strength of team using Simple rating system

LBL: Betting Point Spread

4.2 Algorithm

The proposed model uses a linear Multiple Regression approach adopted from the existing system and K-Nearest Neighbour.

Linear Regression

Input

Attributes X1, X2…. Xn

Main process algorithm

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \varepsilon$

β0 = intercept

β1βρ = regression coefficients

ε = residual standard deviation

Output

Y= Dependent variable

K-Nearest Neighbour

Main process

Given a query instance Xq to be classified,

Let x1, x2….xk denote the k instances from training examples that are nearest to Xq.

Return the class that represents the maximum of the k* instances.

i Associate weights with the attributes

ii Assign weights according to the relevance of attributes

iii Assign random weights

iv Calculate the classification error and adjust the weights according to the error

v Repeat till acceptable level of accuracy is reached

Standard Euclidean Distance

$d(xi,xJ) = \sqrt{(\text{For all attributes a} \sum (xi,a - xJ,a)2)}$

Output

Predicted results for weeks 16 and 17

5. RESULT

Predictions were made using both prediction sets and were tested for weeks 16 and 17 of the 2013 NFL season. In both cases, the seasonal moving average of the prediction set, proved to be very effective in predicting the outcome of the games.

The following results were obtained as displayed in the figures.

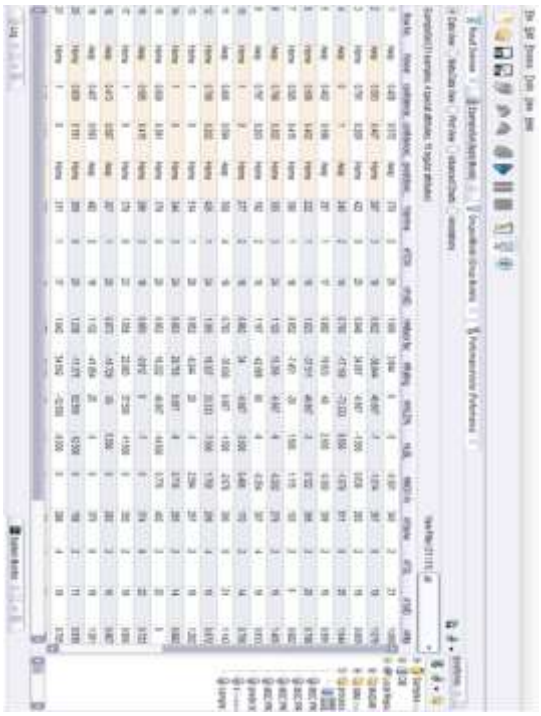

Figure 4. Predicted Results

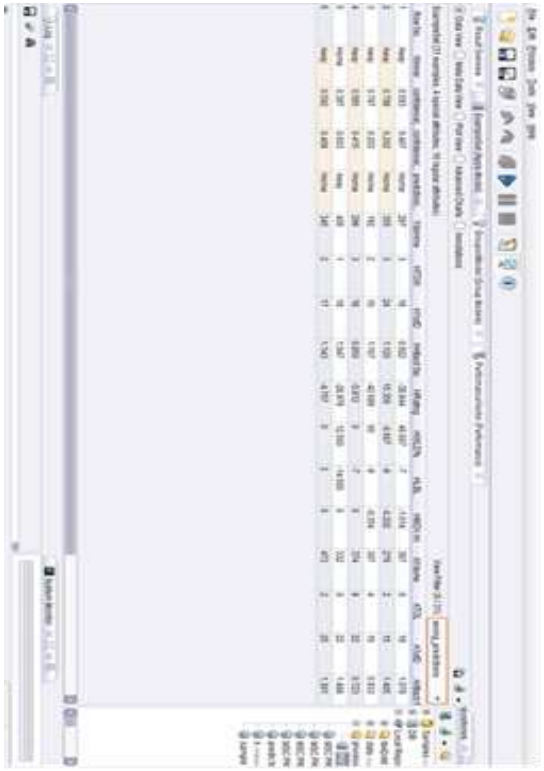

Figure 5. Wrongly Predicted Results

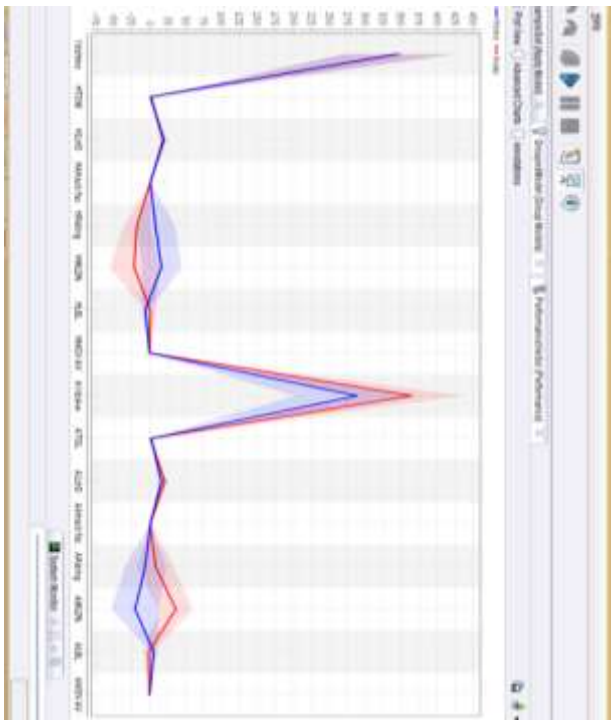

Figure 5. Graph showing predicted Results

6. DISCUSSION OF RESULTS

Prediction sets and were tested for weeks 16 and 17 of the 2013 NFL. Looking at the thirty one games that were predicted, the six games that were incorrectly predicted by the hybrid model over the progression of weeks 16 and 17, two games could be considered "upsets." An upset is when a team defeats a team with a substantially higher winning percentage in a particular game. Three of the remaining size games were games that were "too close to call," or games in which the winner is very difficult to determine. The last game is considered misclassification error in which the hybrid model predicted the incorrect outcome. The prediction accuracy of 80.65% obtained shows that the experiment is substantially better than many existing systems with accuracies of 59.4%, 60.7%, 65.05% and 67.08%.

7. CONCLUSION

The development of a hybrid model using Linear Regression and K- Nearest Neighbour techniques in the prediction of the results of NFL games with an improved accuracy.

The research highlights of this paper are:

- This paper proposes a better approach for sports prediction with unique features.
- The approach uses hybridized data mining techniques.
- The hybridized techniques used are Linear Regression and K- Nearest Neigbour.
- The results show improved prediction accuracy of 80.65%

This Research work can be considered as a successful exploration of using data mining techniques for sports result prediction and it provides a good backbone for future research works.

8. REFERENCES

[1] Borghesi R. 2007. The Home Team weather advantage and biases in the NFL betting market. *Journal of Economics and Business*, 59, 340-354

[2] Paul K. and Joel S. 2006. A Logistic Regression/Markov Chain Model For NCAA Basketball. *Journal of Naval Research Logistics*, vol 53, 1-23

[3] Robert P.S. and Hsinchun C. 2010. A Quantitative Stock Prediction System based on Financial News. Retrieved14/07/2014: http://www.robschumaker.com/publications/IPM%20-%20A%20Quantitative%20Stock%20Prediction%20System%20based%20on%20Financial%20News.pdf

[4] Harville D., (2003). The Selection or Seeding of College Basketball or Football Teams for postseason Competition. Journal of the American Statistical Association, Vol 98, 17-27

[5] Brady T. and Madhur L. 2008. New Application of Linear Modeling in the Prediction of College Football Bowl Outcomes and the Development of Team Ratings. Journal of Quantitative Analysis in Sports, 1-21

[6] Brassington G.B., Freeman J., Huang X., Pugh T., Oke P.R., Sandery P.A., Taylor A., Andreu-Burillo I., Schiller A., Griffin D.A., Fiedler R., Mansbridge J., Beggs H. & Spillman C.M. 2012. Ocean Model, Analysis and Prediction System: version 2 CAWCR Technical Report. The Centre for Australian Weather and Climate Research. No. 052Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[7] Ming X., Donghai W., Jidong G., Keith B. and Kelvin K.D. 2003. The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. Meteorol Atmos Phys, 082, 139–170. DOI: 10.1007/s00703-001-0595-6.

[8] Yonggan W. 2013. The Prediction Model Research on Objectives and Results of Football Game Based on Regression Method. 2nd International Conference on Management Science and Industrial Engineering (MSIE), pp139-143.

International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 48 - 52, 2015, ISSN:- 2319–8656

CLEARMiner: Mining of Multitemporal Remote Sensing Images

Aneesh Chandran
Department of Computer Science and Engineering
Jyothi Engineering College
Cheruthuruthy, Thrissur, India

Swathy Ramadas
Department of Computer Science and Engineering
Jyothi Engineering College
Cheruthuruthy, Thrissur, India

Abstract: A new unsupervised algorithm, called CLimate and rEmote sensing Association patterns Miner, for mining association patterns on heterogeneous time series from climate and remote sensing data, integrated in a remote sensing information system is developed to improve the monitoring of sugar cane fields. The system, called RemoteAgri, consists of a large database of climate data and low-resolution remote sensing images, an image pre-processing module, a time series extraction module, and time series mining methods. The time series mining method transforms series to symbolic representation in order to identify patterns in a multitemporal satellite images and associate them with patterns in other series within a temporal sliding window. The validation process was achieved with agro climatic data and NOAA-AVHRR images of sugar cane fields. Rules generated by the new algorithm show the association patterns in different periods of time in each time series, pointing to a time delay between the occurrences of patterns in the series analyzed, corroborating what specialists usually forecast without having the burden of dealing with many data charts. This new method can be used by agro meteorologists to mine and discover knowledge from their long time series of past and forecasting data, being a valuable tool to support their decision-making process.

Keywords: NOAA-AVHRR Images, Association Rules, Maximum Cross Correlation, Time Series Mining, Sequential Patterns

1. INTRODUCTION

The knowledge discovery, information mining methods and advances in computer technology have contributed to increase the access and application of remote sensing imagery. New technologies developed to be applied in the remote sensing area have increased its use in real applications. However, several users still have problems to deal with satellite images due to different and more sophisticated demands being imposed to them, as well as the fast growing in quantity and complexity of remote sensing data [1]. The knowledge discovery approach has been considered a promising alternative to explore and find relevant information on this huge volume of data. Some initiatives involving information and image mining have been accomplished through different techniques with reasonable results [2]–[4].

Association rules were proposed by Agrawall *et al.* [5] to solve the problem of discovering which items are bought together in a transaction. The number of rules discovered can be so large that analyzing the entire set and finding the most interesting ones can be a difficult task for the user. Then, Klemettinem *et al.* [9] proposed a method based on rule templates to identify interesting rules.

Instead of extracting features from images, other approaches work on computing measurements (indexes) from images generated by a combination of remote sensor channels that can be used to identify the green biomass, and soil temperature, for example. Thus, these indexes (measurements) can be extracted considering each pixel of multitemporal image data sets generating different time series. Time series are generated and studied in several areas, and data mining techniques have been developed to analyze them [5]–[7].

Mannila *et al.* [10] proposed a method to episodal sequential data mining that uses all frequent episodes within one sequence. Zaki [11] proposed the use of temporal constraints in transactional sequences. Harms *et al.* [12] defined methods that combine constraints and closure principles with a sliding window approach. Their objective was to find frequent closed episodes in multiple event sequence. In general, several techniques have been proposed to discover sequential patterns in temporal data in the last decade.

1.1 NOAA-AVHRR

The AVHRR is a radiation-detection imager that can be used for remotely determining cloud cover and the surface temperature. Note that the term surface can mean the surface of the Earth, the upper surfaces of clouds, or the surface of a body of water. This scanning radiometer uses 6 detectors that collect different bands of radiation wavelengths. The first AVHRR was a 4-channel radiometer, first carried on TIROS-N (launched October 1978). This was subsequently improved to a 5-channel instrument (AVHRR/2) that was initially carried on NOAA-7 (launched June 1981). The latest instrument version is AVHRR/3, with 6 channels, first carried on NOAA-15 launched in May 1998. The AVHRR/3 instrument weighs approximately 72 pounds, measures 11.5 inches X 14.4 inches X 31.4 inches, and consumes 28.5 watts power.
Measuring the same view, the array of diverse wavelengths, after processing, permits multi spectral analysis for more precisely defining hydrologic, oceanographic, and meteorological parameters. Comparison of data from two channels is often used to observe features or measure various environmental parameters. The three channels operating entirely within the infrared band are used to detect the heat radiation from and hence, the temperature of land, water, sea surfaces, and the clouds above them.

2. CLimate and rEmote sensing Association patteRns Miner (CLEARMiner)

A new unsupervised algorithm for mining association patterns on heterogeneous time series integrated to a remote sensing information system. The time series mining module was developed to generate rules considering a time lag. To do so, we define the constraint of time window to find association patterns that are extracted in two steps. First, the algorithm transforms multiple time series in a representation of patterns (peaks, mountains, and plateaus), with discrete intervals that maintain the time occurrence and represent phenomena on climate or remote sensing time series. In a second step, the algorithm generates rules that associate patterns in multiple time series with qualitative information. This algorithm-CLimate and rEmote sensing Association patteRns

Miner (CLEARMiner)-uses a sliding window value to find the rules that correspond to the number of patterns by window.

The algorithm quality is assessed using time series of agro meteorological data and multitemporal images from an important region of sugarcane production fields in Brazil. Sugarcane crops have expanded due to different reasons, such as, biofuel production, potential benefits to the environment as a possible way of mitigation of greenhouse gases emission, economic impact, among others. Although traditional ways to assess the sugar cane expansion exist, remote sensing images have been widely adopted to evaluate the direct land conversion to sugar cane. As sugarcane crops are cultivated on large fields, researchers have used satellites of medium and low spatial resolution, such as NOAA-AVHRR, 1 to identify areas for sugarcane expansion. We have also applied CLEARMiner to El Nino time series in order to discover their influence over precipitation distribution regime in regions of South America. In fact, both case studies are suitable to test the CLEARMiner algorithm since both experiments presuppose a relationship between series considering a time lag.

This algorithm works on multiple time series of continuous data, identifying patterns according to a given relevance factor (r) and a plateau length (l) thresholds. In its last step, the algorithm associates patterns according to a temporal sliding window that corresponds to the number of patterns. The number of patterns decreases when the tuning parameters increase, as the experiments showed. Patterns can be seen as discrete intervals that allow the association between series. CLEARMiner presents rules in two formats: short and extended. Short rules are easier to understand, but they are not sufficient to visualize the peak amplitudes and the length of the plateaus. Therefore, the algorithm also presents rules in extended format including details of the values variation and time intervals.

3. ARCHITECTURE OF REMOTE AGRI

Before applying data mining techniques in remote sensing imagery, it is necessary to submit images to the preprocessing process. The knowledge discovery process in information mining systems involves three main phases: data preparation, data mining, and presentation of knowledge. Geometric correction combines indirect navigation and spacecraft attitude error estimation. After that, the maximum cross correlation technique can be used to detect the geographic displacement between the base image and the target one.

Module 1 corresponds to the image georeferencing step executed in batch mode by NAVPro; Module 2 is executed by SatImagExplorer which was proposed to extract values or compute indexes from multitemporal images generating time series for each pixel of the image; and Module 3 refers to time series mining module (CLEARMiner) developed to associate climate data with indexes extracted from NOAA-AVHRR images.

3.1 System Prototype

The system prototype consists of three major components as shown in Fig -1:

- image georeferencing module
- time series extraction module
- time series mining method

The first module to be executed in the RemoteAgri system corresponds to the image georeferencing process, as is presented in figure 2.1. This module is composed of several Cshell scripts that call the subroutines of NAV system in batch mode to accomplish necessary tasks, such as:

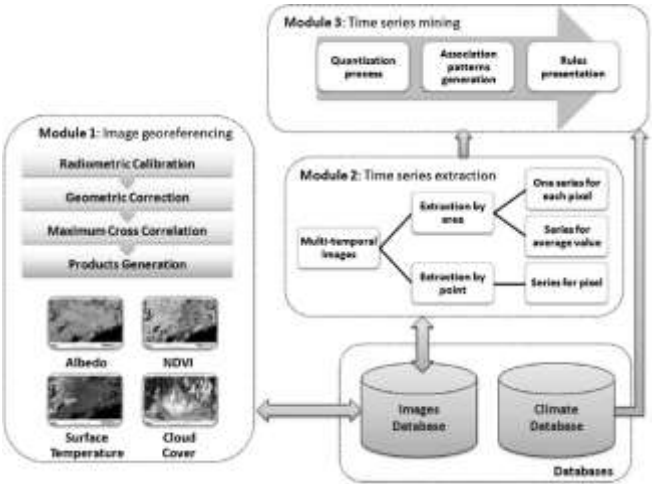

Fig -1: Schematic diagram of the multitemporal images mining system—RemoteAgri

- conversion from raw format to an intermediary;
- radiometric calibration;
- geometric correction;
- identification of pixels classified as cloud.

The georeferencing module allows users to generate four different synthesis images: albedo, NDVI, surface temperature, and cloud cover for a specific region as shown in Fig. 1.

As the volume of images is huge, an extraction module called SatImagExplorer was proposed to perform it faster and in a more flexibly way. The second module extracts values or computes the index from the images opened. Then, it generates a time series computing the index values for all images using the same coordinate (latitude/longitude) of the region. In addition to the direct interaction with the system interface, users can also extract time series using a vector of coordinates that defines the desired region. Time series extracted from multitemporal images SatImagExplorer are then mined in order to discover patterns or association patterns. The last module refers to time series mining developed to associate climate data to indexes extracted from NOAAAVHRR images.

3.2 Maximum Cross Correlation Method

The maximum Cross Correlation (MCC) method is used to automatically compute the satellite attitude parameters required to geometrically correct images to this base image. The MCC method detects the geographic displacements between the base image and the target image[5]. These image displacements are then used to compute the roll, pitch, and yaw attitude parameters. This approach requires the base image to have minimal cloud cover to maximize the potential sites for the computation of image offsets. In the actual application of the base image to the navigation of subsequent images, several levels of cloud detection are applied to ensure that clouds do not influence the image correction calculations.

The base image must be registered to the exact same grid as the target images and must be as cloud free as possible. A second requirement is that the radiance distributions of the base and target images be similar. The widely varying illumination conditions that exist for different orbits prevent the use of the reflected channels for this algorithm.

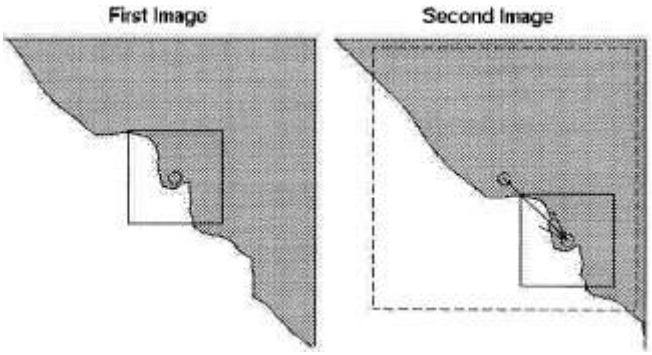

Fig -2: MCC Method applied to two sequential satellite images

4. QUANTZATION PROCESS

The process of time series mining is divided into three parts as shown in figure 2.1. The Quantization process module receives as input a set of remote sensing and climate time series. The time series is defined as a sequence of pairs (ai, ti) with i = 1. . . n, i.e., S = [(a1, t1), . . . , (ai, ti), . . . , (an, tn)] and (t1 < . . . < ti < . . . <tn), where each ai is a data value, and each ti is a time value in which ai occurs. Each pair (ai, ti) is called an evente. A set of events E contains n events of type ei = (ai, ti) for i = 1. . . n. Each ai is a continuous value. Each ti is a unit of time that can be given in days, months or years. Given two sequences S1 and S2, the values ti of both sequences must be measured in the same time unit.

A set of consecutive ei, i.e., Se = (ei, ei+1. . . ek), where ei = (ai, ti) for ti >= t1 and tk <= tn is called the event sequence Se. The number of elements ei in the event sequence depends on the difference between events given by di = (ai+1 - ai) (1st step in figure 3.2), and a given d parameter whose default value is set by the algorithm.

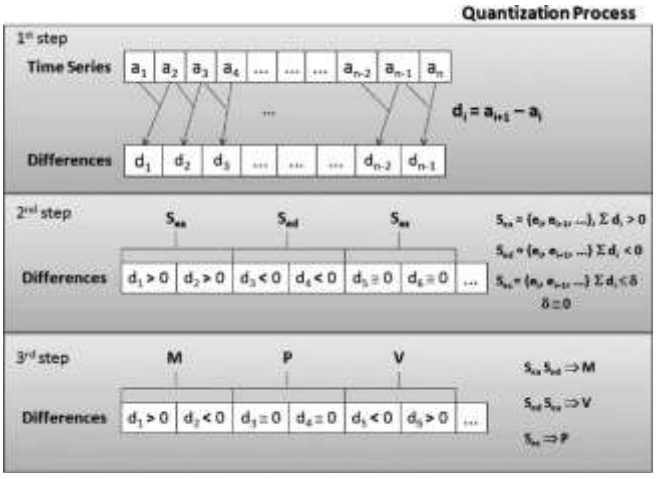

Fig -3: Representation of the three steps of the quantization process executed by time series mining module (CLEARMiner).

V (valley) corresponds to a pattern defined as the concatenation of a descending event sequence and an ascending event sequence (i.e., V = $S_{ed}S_{ea}$). P (plateau) represents a kind of pattern described as a stable event sequence (i.e., P = S_{es}), while M (mountain) indicates a pattern generated by the concatenation of an ascending event sequence and a descending event sequence (i.e., M= $S_{ea}S_{ed}$). Figure 3.3 presents an example of a pattern V. In real data, V can be observed when a sharp drop in the minimum temperature occurs. Algorithm 1 presents the main idea used to convert a time series in a pattern sequence of V, P, and M.

Algorithm 1 PatternsFind Method

Input: Time series S_i; thresholds δ, ρ, λ
Output: Patterns V, M, P
1: **for** $i := 1$ to n
2: calculate the array of differences $d_i = a_{i+1} - a_i$
3: **end for**
4: **for all** d_i values **do**
5: Find S_{ea} = Set of ascending event sequences
6: Find S_{ed} = Set of descending event sequences
7: Find S_{es} = Set of stable event sequences
8: **end for**
9: Eliminate S_{ea} and S_{ed} when $\sum d_i < \rho$
10: Eliminate S_{es} when $\sum d_i < \lambda$
11: **for all** S_e not eliminated **do**
12: V = concatenation of $S_{ed}S_{ea}$
13: M = concatenation of $S_{ea}S_{ed}$
14: $P = S_{es}$
15: **end for**
16: Set of all patterns as $[a_{init}, a_i, a_{end}](t_{init}, t_{end})$

The algorithm concatenates consecutive sequences Sea and Sed to generate an M pattern, Sed and Sea to generate a V pattern, and Ses to generate P patterns.

5. ASSOCIATION PATTERNS GENERATION

A pattern in one time series can be associated to patterns in other time series. Consider an association pattern as an expression of the form Si[a] => Sj [b], where Si and Sj are different time series (for example, rainfall series), a and b are frequent patterns, such as M, V , or P.

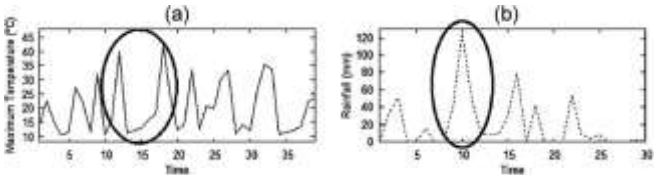

Fig -4: (a) Pattern of type V is similar to negative peaks. (b)Pattern of type M is equivalent to a positive peak.

The support of Si[a] => Sj [b] represents the frequency of occurrences and is given by

$$support = \frac{fr\{S_i[\alpha], S_j[\beta]\})}{T}$$

The confidence measure indicates the percentage of all patterns in Si and Sj containing Si[a] that also contain Sj [b]. The confidence for the rule Si[a]) Sj [b] is given by,

$$conf = \frac{fr\,(S_i[\alpha], S_j[\beta])}{fr\,(S_i[\alpha])}$$

Algorithm 2 shows the pseudocode for CLEARMiner. The CLEARMiner algorithm calculates j-frequentPatterns for each time series. The algorithm only stores j-frequentPatterns greater than the *min_sup* threshold defined by the user.

Algorithm 2 CLEARMiner Algorithm

Input: Dataset A of k time series structured as $\{e_1, e_2, \ldots, e_n\}$ where e_i is an event of time series S_i and k is the number of time series; p is frequent pattern and m is the number of patterns; thresholds δ, ρ, λ and w
Output: The mined rules
Scan data set A
2: **for** each time series S_i **do**
PatternsFind$(S_i, \delta, \rho, \lambda)$
4: **end for**
$F_1 = \{1 - \text{frequentPattern } (S_i[\langle pattern \rangle])\}$
6: **for** $p = 2; p \leq m; p = p + 1$ **do**
$C_p =$ Set of candidate p-frequentPattern
8: $(S_i[\langle pattern \rangle]S_j[\langle pattern \rangle]$ and so on)
for all input-frequentPatterns in the data set **do**
10: increment count of all p-frequentPattern $\in C_l$
end for
12: $F_p = \{frequentPattern \in C_p |$
$sup(frequentPattern) \geq min\_sup\}$
14: **end for**
for all w **do**
16: RuleGenerate(F_p, min_conf)
end for

For each frequent pattern in F, the algorithm calculates, via the RuleGenerate Method, the confidence value (line 2- Algorithm 3). If confidence is greater than minconf, it generates rules (lines 3 to 5-Algorithm 3).

Algorithm 3 RuleGenerate Method

Input: F_p and min_conf
Output: The mined rules
for all frequentPattern $S_i[\alpha]$ and $S_j[\beta] \in F_p$ **do**
$conf = fr(S_i[\alpha], S_j[\beta])/fr(S_i[\alpha])$
3: **if** $conf \geq min\_conf$ **then**
output the rule $S_i[\alpha] \Rightarrow S_j[\beta]$ and $conf$
end if
6: **end for**

6. RULES PRESENTATION

The algorithm presents the rules in two formats to better visualize them: short (the succinct way) and extended (those with more details and time stamp)[10]. The short format is more succinct and easier to be analyzed. However, it contains no information about the context in which the phenomenon occurred.

This rule indicates that the pattern [ai, ak, an] occurred in the period (tinit1 - tend1) for the time series S1, which is associated to the pattern [aj, al, am] occurred in the period (tinit2 - tend2) for the series S2 with tinit1 tinit2 and tend1 tend2. Thus, the user can analyze rules in the short format to verify correlations between time series and to use the extended format to obtain more details. An example with real data is presented in Fig 3.

7. ADVANTAGES

The results are shown by comparing the CLEARMiner algorithm with two classical and baseline algorithms, apriori [7] and the generalized sequential pattern (GSP) algorithm.
Both algorithms were performed in the Weka platform. As the two algorithms work with discrete data, we compared only the rules generation. The data sets used to run apriori and GSP were quantized by CLEARMiner to avoid distortions that could be

Fig -5: Examples of rules in short and extended format in time series mining module.

caused by different quantization processes. The apriori algorithm mined few rules and did not consider time of occurrences.

The GSP algorithm scans the database several times to generate a set of candidate k-sequences and to calculate their support. We executed the GSP algorithm with min_sup =0.2. For minsup values above 0.2, the GSP algorithm in Weka did not work properly. The sequences mined by GSP are similar to the rules generated by CLEARMiner. However, both algorithms (Apriori and GSP) do not keep information about the time occurrence of the events.
CLEARMiner generates rules in an extended format, which can be used to obtain more details about the correlation between time series.

Another advantage of this method is the quantization process that is executed as a first step. This quantization generates a representation that encompasses the semantics meaningful for climate and agroclimate time series. The criteria to quantize time series are based on phenomena that are observed by meteorologists and agrometeorologists and impact the environment.

8. APPLICATIONS

The multitemporal NDVI images from NOAA-AVHRR were studied, covering the scene with orbit/point 220/75 of Landsat satellite. We have selected regions located in Sao Paulo, which is responsible for the majority of sugar cane production in the country. Sugar cane crops are cultivated in plain relief. The climate of this region presents fluctuations in temperature during the rainy season: October to March. The results of experiments were performed on two real data sets to evaluate and validate the proposed algorithm. The results from such experiments followed the specialists' expectations and helped on tuning the algorithms' parameters. Table I presents a summary of the data sets used, giving their dimensions number (E) and the size of time series (N).

Table -1: Definition of datasets that was used to evaluate the performance of CLEARMiner

Name	Description	E	N
Sugar Cane	Real data composed of NDVI and WRSI values token from the 5 sugar cane productive areas Sao Paulo State (Brazil) from 04/01/2001 to 03/31/2008	2	≅ 500
El Nino	Real data composed of temperature and anomalies for 4 regions in the Pacific Ocean and rainfall of Quarai, Brazil	9	500

Two main applications are:

- Mining NDVI and WRSI Time Series From Sugar Cane Regions
- Mining Time Series of Rainfall and Anomalies Related to El Nino

9. CONCLUSION

The system, called RemoteAgri, consists of a large database of climate data and low-resolution remote sensing images, an image preprocessing module, a time series extraction module, and time series mining methods [1]. The preprocessing module was projected to perform accurate geometric correction, what is a requirement particularly for land and agriculture applications of satellite images. The time series extraction is accomplished through a graphical interface that allows easy interaction and high flexibility to users. The time series mining method transforms series to symbolic representation in order to identify patterns in a multitemporal satellite images and associate them with patterns in other series within a temporal sliding window.

The results show that the algorithm detects some association patterns that are known by experts, as expected, indicating the correctness and feasibility of the proposed method. Moreover, other patterns detected using the highest relevance factors are coincident with extreme phenomena as many days without rain or heavy rain as the specialists suppose to. The mined rules for the relevance patterns indicate a relation between series, allowing these patterns (phenomena) happen in different intervals of time. This method can be used by agrometeorologists to mine and discover knowledge from their long time series of past and forecasting data, being a valuable tool to support their decision-making process.

10. REFERENCES

[1] Luciana Alvim S. Romani, Ana Maria H. de Avila "A New Time Series Mining Approach Applied to Multitemporal Remote Sensing Imagery", Communications of the ACM, 21:140-150, 2013.
[2] R.M. Datcu, H. Daschiel, A. Pelizzari, M. Quartulli. "Information mining in remote sensing image archives: System concepts", 2003.
[3] J. Li and R. M. Narayanan "Integrated spectral and spatial information mining in remote sensing imagery", Communications of the ACM, 54(8):62–71, 2011.
[4] Witold Pedrycz G. B. Yingxu Wang. "Information mining in remote sensing image archives: System evaluation", IEEE Trans. Geosci. Remote Sens, page 188-199, 2005.
[5] W. Emery, D. G. Baldwin, and D.Matthews, "Maximum cross correlation automatic satellite image navigation and attitude corrections for open ocean image navigation", IEEE Proceedings.
[6] J. C. D. M. Esquerdo, J. F. G. Antunes, D. G. Baldwin., "An automatic system for AVHRR land surface product generation," 2003.
[7] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases," in Proc. 4th Int. CFDOA, Chicago, IL, 1993, pp. 69-84.
[8] G. Das, K. Lin, H. Mannila, G. Renganathan, and P. Smyth, "Rule discovery from time series," in Proc. 4th ICKDDM, New York, 1998, pp. 16-22.
[9] R. Srikant and R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements," in Proc. ICEDT, Avignon, France, 1996, pp. 3-17.
[10] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkano, "Finding interesting rules from large sets of discovered association rules," in Proc. CIKM, Gaitherburg, MD, 1994, pp. 401-407.
[11] H. Mannila, H. Toivonem, and A. I. Verkamo, "Discovery of frequent episodes in event sequences," Data Mining Knowl. Discovery, vol. 1, no. 3, pp. 259-289, 1997.
[12] M. Zaki, "Sequence mining in categorical domains: Incorporating constraints," in Proc. CIKM, Washington, DC, 2000 pp. 422-429.
[13] S. K. Harms, J. Deogun, J. Saquer, and T. Tadesse, "Discovering representative episodal association rules from event sequences using frequent closed episode sets and event constraints," in Proc. ICDM, San Jose, CA, 2001,pp. 603-606.
[14] J. Wang and J. Han, "BIDE: Efficient mining of frequent closed sequences," in Proc. ICDE, 2004, pp. 79-90

International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 53 - 57, 2015, ISSN:- 2319–8656

Designing of Semantic Nearest Neighbor Search: Survey

Pawar Anita R.
Student BE Computer
S.B. Patil College of Engineering
Indapur, Pune,Maharashtra
India

Pansare Rajashree B.
Student BE Computer
S.B. Patil College of Engineering
Indapur, Pune, Maharashtra
India.

Mulani Tabssum H.
Student BE Computer
S.B. Patil College of Engineering
Indapur, Pune, Maharashtra
India

Bandgar Shrimant B.
Asst. Professor
S.B. Patil College of Engineering
Indapur, Pune, Maharashtra
India

Abstract: Conventional spatial queries, such as range search and nearest neighbor retrieval, involve only conditions on objects' geometric properties. Today, many modern applications call for novel forms of queries that aim to find objects satisfying both a spatial predicate, and a predicate on their associated texts. For example, instead of considering all the restaurants, a nearest neighbor query would instead ask for the restaurant that is the closest among those whose menus contain "steak, spaghetti, brandy" all at the same time. Currently the best solution to such queries is based on the IR2-tree, which, as shown in this paper, has a few deficiencies that seriously impact its efficiency. Motivated by this, we develop a new access method called the spatial inverted index that extends the conventional inverted index to cope with multidimensional data, and comes with algorithms that can answer nearest neighbor queries with keywords in real time. As verified by experiments, the proposed techniques outperform the IR2-tree in query response time significantly, often by a factor of orders of magnitude.

Keywords: spatial Index, K-mean, Merge multiple, keyword-based Apriori item-set, neighbour search

1. INTRODUCTION

1.1 Concept of spatial index

A spatial database manages multidimensional objects(such as points, rectangles, etc.), and provides fast access to those objects based on different selection criteria .The importance of spatial databases is reflected by the convenience of modelling entities of reality in a geometric manner[5][6][7][8]. For example, locations of restaurants, hotels, hospitals and so on are often represented as points in a map, while larger extents such as parks, lakes, and landscapes often as a combination of rectangles. Many functionalities of a spatial database are useful in various ways in specific contexts. For instance, in a geography information system, range search can be deployed to find all restaurants in a certain area; while nearest neighbour retrieval can discover the restaurant closest to a given address.

Today, the widespread use of search engines has made it realistic to write spatial queries in a brand new way. Conventionally, queries focus on objects' geometric properties only, such as whether a point is in a rectangle, or how close two points are from each other. We have seen some modern applications that call for the ability to select objects based on both of their geometric coordinates and their associated texts. For example, it would be fairly useful if a search engine can be used to find the nearest restaurant that offers "steak, spaghetti, and brandy" all at the same time. Note that this is not the "globally" nearest restaurant (which would have been returned by a traditional nearest neighbour query), but the nearest restaurant among only those providing all the demanded foods and drinks. There are easy ways to support queries that combine spatial and text features. For example, for the above query, we could first fetch all the restaurants whose menus contain the set of keywords {steak, spaghetti, brandy}, and then from the retrieved restaurants, find the nearest one. Similarly, one could also do it reversely by targeting first the spatial conditions – browse all the restaurants in ascending order of their distances to the query point until encountering one whose menu has all the keywords.

The major drawback of these straightforward approaches is that they will fail to provide real time answers on difficult inputs. A typical example is that the real nearest neighbour lies quite far away from the query point, while all the closer neighbours are missing at least one of the query keywords.

1.2. Concept of IR2-tree.

Spatial queries with keywords have not been extensively explored. In the past years, the community has sparked enthusiasm in studying keyword search in relational databases[1]. It is until recently that attention was diverted to multidimensional data. The best method to date for nearest neighbour search with keywords is due to Felipe et al. They nicely integrate two well-known concepts: R-tree, a popular spatial index, and signature file, an effective method for keyword-based document retrieval. By doing so they develop a structure called the IR2-tree, which has the strengths of both

R-trees and signature files. Like R-trees, the IR2-tree preserves objects' spatial proximity, which is the key to solving spatial queries efficiently[2]. On the other hand, like signature files, the IR2-tree is able to filter a considerable portion of the objects that do not contain all the query keywords, thus significantly reducing the number of objects to be examined. The IR2-tree, however, also inherits a drawback of signature files: false hits[2]. That is, a signature file, due to its conservative nature, may still direct the search to some objects, even though they do not have all the keywords. The penalty thus caused is the need to verify an object whose satisfying a query or not cannot be resolved using only its signature, but requires loading its full text description, which is expensive due to the resulting random accesses. It is noteworthy that the false hit problem is not specific only to signature files, but also exists in other methods for approximate set membership tests with compact storage. Therefore, the problem cannot be remedied by simply replacing signature file with any of those methods.

1.3 Concept of merge multiple

we design a variant of inverted index that is optimized for multidimensional points, and is thus named the spatial inverted index(SI-index). This access method successfully incorporates point coordinates into a conventional inverted index with small extra space, owing to a delicate compact storage scheme. Mean while, an SI-index preserves the spatial locality of data points, and comes with an R-tree built on every inverted list at little space overhead. As a result, it offers two competing ways for query processing. We can (sequentially) merge multiple lists very much like merging traditional inverted lists by ids. Alternatively, we can also leverage the R-trees to browse the points of all relevant lists in ascending order of their distances to the query point. As demonstrated by experiments, the SI-index significantly outperforms the IR2-tree in query efficiency, often by a factor of orders of magnitude[2].

1.4 Concept of keyword-based nearest neighbour search

There are many process mining algorithms and representations, making it difficult to choose which algorithm to use or compare results. Process mining is essentially a machine learning task, but little work has been done on systematically analyze in algorithms to understand their fundamental properties, such a show much data are needed for confidence in mining. We propose a framework for analyzing process mining algorithms. Processes are viewed as distributions over traces of activities and mining algorithms as learning these distributions. The access to a large quantity of textual documents turns out to be effectual because of the growth of the digital libraries, web, technical documentation, medical data and more. These textual data comprise of resources which can be utilized in a better way. Text mining is major research field due to the need of acquiring knowledge from the large number of available text documents, particularly on the web. Both text mining and data mining are part of information mining and identical in some perspective. Text mining can be described as a knowledge intensive process in which a user communicates with a collection of documents. In order to mine large document collections, it is require pre-processing the text documents and saving the data in the data structure, which is suitable for processing it further than a plain text file. Information Extraction is defined as the mapping of natural language texts like text database, WWW pages, electronic mail etc. into predefined structured representation, or templates which, when filled, represent an extract of key information from the original text.

2. Problem Definition

Let P be a set of multidimensional points. As our goal is to combine keyword search with the existing location-finding services on facilities such as hospitals, restaurants, hotels, etc., we will focus on dimensionality, but our technique can be extended to arbitrary dimensionalities with no technical obstacle .We will assume that the points in P have integer coordinates, such that each coordinate ranges in [0, t], where t is a large integer. This is not as restrictive as it may seem, because even if one would like to insist on real-valued coordinates, the set of different coordinates represent able under a space limit is still finite and enumerable; therefore, we could as well convert everything to integers with proper scaling. As with, each point p 2 P is associated with a set of words, which is denoted as Wp and termed the document of p. For example, if p stands for a restaurant, Wp can be its menu, or if p is a hotel, Wp can be the description of its services and facilities, or if p is a hospital, Wp can be the list of its out-patient specialities. It is clear that Wp may potentially contain numerous words. Traditional nearest neighbour search returns the data point closest to a query point. Following, we extend the problem to include predicates on objects' texts. Formally, in our context, a nearest neighbour (NN) query specifies a point q and a set Wq of keywords (we refer to Wq as the document of the query). It returns the point in Pq that is the nearest to q, where Pq is defined as,

$$Pq=\{p2P|Wq\_Wp\} \quad (1)$$

In other words, Pq is the set of objects in P whose documents contain all the keywords in Wq. In the case where Pq is empty, the query returns nothing. The problem definition can be generalized to k nearest neighbour (kNN) search, which finds the k points in Pq closest to q; if Pq has less than k points, the entire Pq should be returned. For example, assume that P consists of 8 points whose locations are as shown in Figure 1a (the black dots), and their documents are given in Figure 1b. Consider a query point q at the white dot of Figure 1a with the set of keywords

$$Wq=\{c,d\} \quad (2)$$

Nearest neighbour search finds p6, noticing that all points closer to q than p6 are missing either the query keyword c or d. If k = 2 nearest neighbours are wanted, p8 is also returned in addition. The result is still {p6, p8} even if k increases to 3 or higher, because only 2 objects have the keywords c and d at the same time. We consider that the dataset does not fit in memory, and needs to be indexed by efficient access methods in order to minimize the number of I/Os in answering a query.

A spatial database manages multidimensional objects (such as points, rectangles, etc.), and provides fast access to those objects based on different selection criteria. The importance of spatial databases is reflected by the convenience of modeling entities of reality in a geometric manner[5][6][7][8]. For example, locations of restaurants, hotels, hospitals and so on are often represented as points in a map, while larger extents such as parks, lakes, and landscapes often as a combination of rectangles. Many functionalities of a spatial database are useful in various ways in specific contexts. For instance, in a geography information system, range search can be deployed to find all restaurants in a certain area, while nearest neighbor retrieval can discover the

restaurant closest to a given address Today, the widespread use of search engines has made it realistic to write spatial queries in a brand new way.

Conventionally, queries focus on objects' geometric properties only, such as whether a point is in a rectangle, or how close two points are from each other. We have seen some modern applications that call for the ability to select objects based on both of their geometric coordinates and their associated texts. For example, it would be fairly useful if a search engine can be used to find the nearest restaurant that offers "steak, spaghetti, and brandy" all at the same time. Note that this is not the "globally" nearest restaurant (which would have been returned by a traditional nearest neighbor query), but the nearest restaurant among only those providing all the demanded foods and drinks.

There are easy ways to support queries that combine spatial and text features. For example, for the above query, we could first fetch all the restaurants whose menus contain the set of keywords {steak, spaghetti, brandy}, and then from the retrieved restaurants, find the nearest one. Similarly, one could also do it reversely by targeting first the spatial conditions – browse all the restaurants in ascending order of their distances to the query point until encountering one whose menu has all the keywords. The major drawback of these straightforward approaches is that they will fail to provide real time answers on difficult inputs. A typical example is that the real nearest neighbor lies quite far away from the query point, while all the closer neighbors are missing at least one of the query keywords.

Spatial queries with keywords have not been extensively explored. In the past years, the community has sparked enthusiasm in studying keyword search in relational databases[1]. It is until recently that attention was diverted to multidimensional data.

The best method to date for nearest neighbor search with keywords is due to Felipe et al. They nicely integrate two well known concepts: R-[2], a popular spatial index, and signature file, an effective method for keyword-based document retrieval. By doing so they develop a structure called the IR2-tree, which has the strengths of both R-trees and signature files. Like R-trees, the IR2-tree preserves objects' spatial proximity, which is the key to solving spatial queries efficiently. On the other hand, like signature files, the IR2-tree is able to filter considerable portion of the objects that do not contain all the query keywords, thus significantly reducing the number of objects to be examined.

The IR2-tree, however, also inherits a drawback of signature files: false hits. That is, a signature file, due to its conservative nature, may still direct the search to some objects, even though they do not have all the keywords. The penalty thus caused is the need to verify an object whose satisfying a query or not cannot be resolved using only its signature, but requires loading its full text description, which is expensive due to the resulting random accesses. It is noteworthy that the false hit problem is not specific only to signature files, but also exists in other methods for approximate set membership tests with compact storage. Therefore, the problem cannot be remedied by simply replacing signature file with any of those methods.

Data fusion and multicue data matching are fundamental tasks of high-dimensional data analysis. In this paper, we apply the recently introduced diffusion framework to address these tasks. Our contribution is three-fold: First, we present the Laplace- Beltrami approach for computing density invariant embeddings which are essential for integrating different sources of data. Second, we describe a refinement of the Nystro¨m extension algorithm called "geometric harmonics." We also explain how to use this tool for data assimilation. Finally, we introduce a multicue data matching scheme based on nonlinear spectral graphs alignment. The effectiveness of the presented schemes is validated by applying it to the problems of lip-reading and image sequence alignment..

3. Code Review Technique

There are so many techniques to apply the nearest elementas well as locations**.**

3.1 The k-means algorithm

The k-means algorithm is a simple iterative method to partition a given dataset into a user specified number of clusters, *k*. This algorithm has been discovered by several researchers across different disciplines, most notably Lloyd.

3.2 Support vector machines

In today's machine learning applications, support vector machines (SVM) [83] are considered must try—it offers one of the most robust and accurate methods among all well-known algorithms. It has a sound theoretical foundation, requires only a dozen examples for training, and is insensitive to the number of dimensions. In addition, efficient methods for training SVM are also being developed at a fast pace.

3.3 The Apriori algorithm

One of the most popular data mining approaches is to find frequent item sets from a transaction dataset and derive association rules. Finding frequent itemsets (itemsets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence.

3.4 The EM algorithm

Finite mixture distributions provide a flexible and mathematical-based approach to the modelling and clustering of data observed on random phenomena. We focus here on the use of normal mixture models, which can be used to cluster continuous data and to estimate the underlying density function. These mixture models can be fitted by maximum likelihood via the EM (Expectation–Maximization) algorithm.

3.5 CART

The 1984 monograph, "CART: Classification and Regression Trees," co-authored by Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone, [9] represents a major milestone in the evolution of Artificial Intelligence, Machine Learning, non-parametric statistics, and data mining. The work is important for the comprehensiveness of its study of decision trees, the technical innovations it introduces, its sophisticated discussion of tree structured data analysis, and its authoritative treatment of large sample theory for trees[2]. While CART citations can be found in almost any domain, far more appear in fields such as electrical engineering, biology, medical research and financial topics than, for example, in

marketing research or sociology where other tree methods are more popular. This section is intended to highlight key themes treated in the CART monograph so as to encourage readers to return to the original source for more detail.

4. R-trees System

An SI-index is no more than a compressed version of an ordinary inverted index with coordinates embedded, and hence, can be queried in the same way as described i.e., by merging several inverted lists. In the sequel, we will explore the option of indexing each inverted list with an R-tree[2]. As explained in these trees allow us to process a query by distance browsing, which is efficient when the query keyword set we is small. Our goal is to let each block of an inverted list be directly a leaf node in the R-tree. This is in contrast to the alternative approach of building an R-tree that shares nothing with the inverted list, which wastes space by duplicating each point in the inverted list. Furthermore, our goal is to offer two search strategies simultaneously merging and distance browsing .As before, merging demands those points of all lists should be ordered following the same principle. This is not a problem because our design in the previous subsection has laid down such a principle: ascending order of Z-values. Moreover, this ordering has a crucial property that conventional id-based ordering lacks: preservation of spatial proximity. The property makes it possible to build good R-trees without destroying the Z-value ordering of any list. Specifically, we can (carefully) group consecutive points of a list into MBRs, and incorporate all MBRs into an R-tree[2]. The proximity preserving nature of the Z-curve will ensure that the MBRs are reasonably small when the dimensionality is low.

5. Proposed System

Our treatment of nearest neighbor search falls in the general topic of spatial keyword search, which has also given rise to several alternative problems. A complete survey of all those problems goes beyond the scope of this project. 1. Strictly speaking, this is not precisely true because merging may need to jump across different lists; however, random I/Os will account for only a small fraction of the total overhead as long as a proper perfecting strategy is employed, e.g., reading 10 sequential pages at a time. Considered a form of keyword-based nearest neighbor queries that is similar to our formulation, but differs in how objects' texts play a role in determining the query result. Specifically, aiming at an IR flavor, the approach of computes the relevance between the documents of an object p and a query q. This relevance score is then integrated with the Euclidean distance between p and q to calculate an overall similarity of p to q. The few objects with the highest similarity are returned. In this way, an object may still be in the query result, even though its document does not contain all the query keywords. In our method, same as, object texts are utilized in evaluating a Boolean predicate, i.e., if any query keyword is missing in an object's document, it must not be returned. Neither approach subsumes the other, nor do both make sense in different applications

4.1 System Architecture

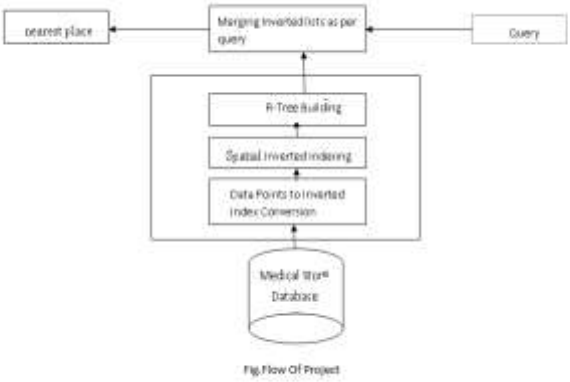

Fig Flow Of Project

As an application in our favor, consider the scenario where we want to find a close restaurant serving "steak, spaghetti and brandy", and do not accept any restaurant that does not serve any of these three items. In this case, a restaurant's document either fully satisfies our requirement, or does not satisfy at all. There is no "partial satisfaction", as is the rationale behind the approach of, In geographic web search, each webpage is assigned a geographic region that is pertinent to the webpage's contents. In web search, such regions are taken into account so that higher rankings are given to the pages in the same area as the location of the computer issuing the query (as can be inferred from the computer's IP address) . The underpinning problem that needs to be solved is different from keyword-based nearest neighbor search, but can be regarded as the combination of keyword search and range queries. Specifically, let P be a set of points each of which carries a single keyword. Given a set Wq of query keywords (note: no query point q is needed), the goal is to find m = |Wq| points from P such that (i) each point has a distinct keyword in Wq, and (ii) the maximum mutual distance of these points is minimized (among all subsets of m points in P fulfilling the previous condition).

In other words, the problem has a "collaborative" nature in that the resulting m points should cover the query keywords together. This is fundamentally different from our work where there is no sense of collaboration at all, and instead the quality of each individual point with respect to a query can be quantified into a concrete value. Proposed collective spatial keyword querying, which is based on similar ideas, but aims at optimizing different objective functions[4][12].

6. Conclusion

This paper describe, A user-set minimum support decides about which rules have high support .Once the rules are selected, they are all treated the same, irrespective of how high or how low their support. Their locations are uniformly distributed in Uniform, whereas in Skew, they follow the Zip f distribution. For both datasets, the vocabulary has 200 words, and each word appears in the text documents of 50kpoints. The difference is that the association of words with points is completely random in Uniform, while in Skew, there is a pattern of "word-locality": points that are spatially close have almost identical text documents.

7. REFERENCES

[1] S. Agrawal , S.Chaudhuri,and G.Das.Dbxplorer:A system for keyword-based search over relational databases. In Proc.Of International Conference on DataEngin-eering (ICDE), pages 516, 2002.Ding, W. and Marchionini, G. 1997 A Study on Video Browsing

Strategies. Technical Report. University of Maryland at College Park.

[2] N.Beckmann, H.Kriegel,R.Schneider, and B.Seeger.The R*-tree:Ane_cient and robust access method for points and rectangles.In Procof ACM Management of Data(SIGMOD), pages 322331, 1990.Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[3] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan. Keyword searching and browsing in databases using banksInProc.of International Conf- erence on Data Engineering (ICDE),pages 431440, 2002.

[4] X .Cao, L .Chen, G.Cong, C. S.Jensen, Q.Qu, A.Skovsgaard,D.Wu, and M.L.Yiu.Spatial keyword querying. In ER, pages 1629, 2012.Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[5] X.Cao, G.Cong, and C.S.Jensen.Retrieving top-k prestige-based relevant spatial web objects. PVLDB, (1):373384, 2010

[6] Y .-Y .Chen,T.Suel, and A. Markowetz.E_cient query processing geographic web search engines. In Proc.of ACM Management of Data(SIGMOD),pages277288, 2006

[7] G .Cong, C. S .Jensen, and D. Wu. E_cient retrieval of the top-k most relevant spatial web objects. PVLDB, (1):337348, 2009.

[8] C .Faloutsos and S .Christodoulakis Signature _les:An access method for docu-6mentsand its analytical performance evaluation.ACM Trans-actions on Infor- mation Systems (TOIS),2(4):267288, 1984.

[9] I.D .Felipe, V.Hristidis, and N.Rishe. Keyword search on spatial databases.In Proc.of International Conference on Data Engineering(ICDE) pages 656665, 2008

[10] R. Hariharan,B. Hore, C. Li, and S.Mehrotra.Processing spatial keyword(SK) in geographic information retrieval (GIR) systems. In Proc. of Scienti_c and Statistical Database Management (SSDBM), 2007.

[11] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi .Collective spatial keyword quer-ying .In Proc. of ACM Management of Data (SIGMOD),pages 373384, 2011.

[12] I.Kamel and C.Faloutsos. Hilbert R-tree : An improved r-tree using fractals. In Proc. of Very Large Data Bases(VLDB), pages 500509, 1994.

[13] B .Chazelle, J .Kilian, R .Rubinfeld, and A.Tal. The bloomier _lter:ane_cient data structure for static support lookup tables. In Proc.of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 3039, 2004.

International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 58 - 62, 2015, ISSN:- 2319–8656

Secure Sharing of Personal Health Records in Cloud Computing using Encryption

Sabna A B
Department of Computer Science and Engineering
Jyothi Engineering College
Cheruthuruthy, Thrissur, India

Harsha T D
Department of Computer Science and Engineering
Jyothi Engineering College
Cheruthuruthy, Thrissur, India

Abstract: The PHR is a tool that you can use to collect, track and share past and current information about your health or the health of someone in your care. Personal health record (PHR) is considered as an emerging patient-centric model of health information exchange, where people can share their health information to other people. Since there are wide privacy concerns about the health records and due to high operational cost, users stored at a third party server called as Cloud Server. The issues such as risks of privacy exposure, scalability in key management, access problem, user revocation, have remained the most important challenges towards achieving fine-grained, cryptographically enforced data access control. In order to get rid off from this ,in this paper we introduce attribute-based encryption (ABE) techniques to encrypt each patient's PHR file so that an unauthorised people won't be able to view our PHR file.

Keywords: Personal Health Records, Cloud computing, Attribute Based Encryption.

1. INTRODUCTION

Personal Health Record (PHR) is emerged as a patient- centric model of health information exchange. Nowadays most of the users store their health related data in a third parties on the Internet. It allows the patient to create and control his/her medical related data which may be placed in a single place such as information center. Due to the high cost of building of the sensitive personal health information, especially when they are stored at a third-party server which people may not fully trust example, personal email, data, and personal preferences are stored on web portal sites such as Google and Yahoo. So in this paper we use an encryption called Attribute based Encryption so that people will be able encrypt their PHR file from wherever they want to. The main concern is about the privacy of patients, personal health data and to find which user could gain access to the medical records stored in a cloud server.

In ABE [1], the attributes of users or data that selects the access policies enables a patient to share their PHR selectively among a set of users after encrypting the file on the basis of a set of attributes. As a result, the number of attributes involved determines the complexities in encryption, generation of key and decryption. The Multi Authority Attribute Based Encryption (MA ABE) scheme provides multiple authority based access control mechanism in . The PHR owner should decide how to encrypt their files and how to allow the users to obtain access for each file. A PHR file should only be available to the users who are given the corresponding decryption key, which will be confidential to the rest of users.

By using ABE, to address key management challenges, we divide the users into two types of domains; they are public and personal domain. For personal domain, KP-ABE scheme is used. For public domain, MA-ABE scheme is used and the PHR is under control of outsource agent. Here we propose a novel idea which is an enhance MA-ABE so that, the user will have full control on their own PHR.

Furthermore, the patient will always have the right to not only grant, but also revoke access privileges when the patient feel it is necessary. The main goal of patient-centric privacy is conflict with scalability in PHR system. The authorized users may either want to access PHR file for personal use or professional purposes. Implementation of standards for health-care data, accurate patient identification and matching of records, and definition of incentives for accelerated deployment of health information technology.

2. PROBLEM DEFINITION

In Multi Authority –Attribute Based Encryption the existing key is created by outsourced again the data is endangered so that the key control is visited with the outsource agent and it became difficult to manage. Thus the future enhancement to propose a novel idea which is an enhance MA-ABE so that, key will be given by the user.

3. RELATED WORK

In this paper, most of the related works in cryptographic enforced accessing control for the outsourced data and ABE. To realize fine-grained access control, the traditional public key encryption (PKE)-based schemes [10],[8] either need high key management or require encrypting the multiple copies of a file using different users keys. To improve upon the scalability of the above solutions, one-to-many encryption methods such as ABE can be used. In Goyal et al.'s paper on ABE [11], data's are encrypted based on a set of attributes so that multiple users who possess proper keys can decrypt. This will potentially makes encryption and key management more efficient [12].

Fine grained access control systems facilitate granting differential access rights to a set of users and specify the access rights of individual users. Several techniques are known for implementing the grained access control.

They also note how their techniques for resisting collusion attacks are useful in attribute-based encryption. However, the cost of their scheme in terms of computation, private key size, and cipher text size increases exponentially with the number of attributes. We also note that there has been other work that applied IBE techniques to access control.

4. FRAMEWORK

In this paper, the purpose of our framework is to provide security for patient-centric Personal Health Record access and key management in an efficient manner at the same time[14]. If the users attribute is not valid, then the user will be unable to access the future Personal Health Record files using the attributes. The PHR data should support the users from personal domain as well as public domain. The public domain may have more number of users who may be in huge number and unable to predict, so that the system should be highly scalable in terms of the complexity in key management system communication, computation and storage. The owner in managing users and keys should be minimized to enjoy usability Fig-1 By using the ABE, encryption of personal health records self-protective, that is they can access only authorized users on a semi trusted server.

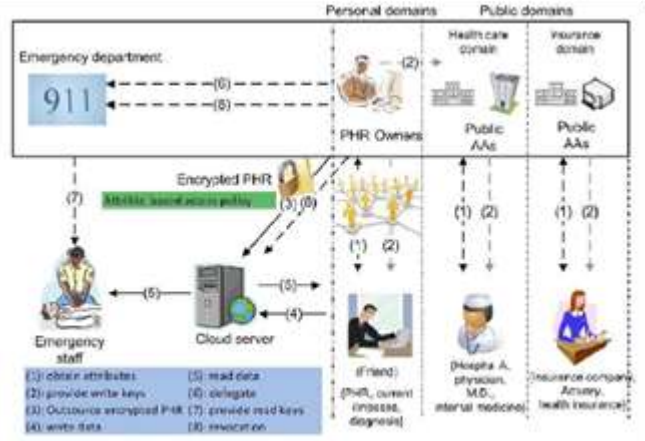

Fig -1: Framework of PHR

5. DESIGN GOALS

In this paper, our main goal is to provide the security for the data files present in the cloud server. Especially we allow each and every data owner to provide the access policy for each data. The users are given with a set of attributes and their corresponding keys. The individual users can only decrypt the files if and only if the corresponding set of attributes matches with the access policy. In addition to that, we handle the users who are revoked. That is users who are unauthorized but once upon a time authorized must not be able to access the data.

In the case of Maintaining Confidentiality, it allows the unauthorized users are not allowed to read data file or modify the data file and thus maintaining the confidentiality of each data file in the cloud server. In Data Access,the data access can be described in two ways[3]. First of all ,any member of the group can access the data present in the cloud. Second, unauthorized and revoked users cannot gain the access to the files of the cloud resources

6. PROPOSED SCHEME

The Personal Health Records are maintained in the data server under the cloud environment. A novel framework for secure and sharing of the personal health records has been proposed in this paper. The Public access and Personal access models are designed with the security and the privacy enabled mechanism Fig-2. The framework addresses the unique challenges brought by the multiple PHR owners and the users, so that the complexity of key management is greatly got reduced. The attribute-based encryption model is enhanced to support the operations with the Multi Authority Attribute Based Encryption. The System will improve its dynamic policy management model. Thus, Personal Health Records are maintained with the security and privacy.

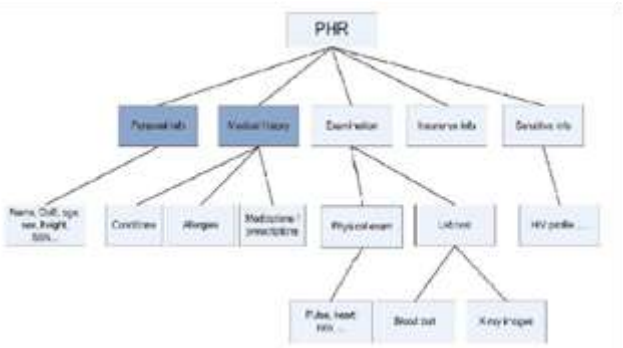

Fig -2: Attribute Hierarchy of files

The solution we propose is based on the following assumptions:

- There is a trusted authority (TA) who generates the keys for the users of the system. There is also a public directory that is used by the TA to publish the system public values (such as public keys) and the parameters that are needed for cryptographic operations.

- A user is associated with a unique identifier (ID), and (ii) a set of attributes (ω). Each user has a public key and a private key. The Private Key is generated and issued by the TA after verification of the user's attributes.

- The health record database is hosted on the cloud storage. The cloud server is trusted for performing the requested operation but will not be able to do other unspecified operations such as reading patients' data. Therefore the health information on the storage must be kept in secured form.

PHR encryption and access. The owners upload ABEencrypted PHR files to the server (3). Each owner's PHR file is encrypted both under a certain fine-grained and rolebased access policy for users from the PUD to access, and under a selected set of data attributes that allows access from users in the PSD. Only authorized users can decrypt the PHR files, excluding the server. For improving efficiency, the data attributes will include all the intermediate file types from a leaf node to the root. For example, in Fig. 2, an "allergy" file'sattributes are fPHR; medical history; allergyg. The data readers download PHR files from the server, and they can decrypt the files only if they have suitable attribute-based keys (5). The data contributors will be granted write access to someone's PHR, if they present proper write keys.

User revocation. Here, we consider revocation of a data reader or her attributes/access privileges[15]. There are several possible cases:

1. revocation of one or more role attributes of a public domain user;

2. revocation of a public domain user which is equivalent to revoking all of that user's attributes. These operations are done by the AA that the user belongs to, where the actual computations can be delegated to the server to improve efficiency (8).

3. Revocation of a personal domain user's access privileges;

4. revocation of a personal domain user. These can be initiated through the PHR owner's client application in a similar way.

Policy updates. A PHR owner can update her sharing policy for an existing PHR document by updating the attributes (or access policy) in the ciphertext. The supported operations include add/delete/modify, which can be done by the server on behalf of the user.

Break-glass. When an emergency happens, the regular access policies may no longer be applicable. To handle this situation, break-glass access is needed to access the victim's PHR. In our framework, each owner's PHR's access right is also delegated to an emergency department (ED, (6)). To prevent from abuse of break-glass option, the emergency staff needs to contact the ED to verify her identity .

Remarks. The separation of PSD/PUD and data/role attributes reflects the real-world situation. First, in the PSD, a patient usually only gives personal access of his/her sensitive PHR to selected users, such as family members and close friends, rather than all the friends in the social network. Different PSD users can be assigned different access privileges based on their relationships with the owner. In this way, patients can exert fine-control over the access for each user in their PSDs. Second, by our multidomain and multiauthority framework, each public user only needs to contact AAs in its own PUD who collaboratively generates a secret key for the user, which reduces the workload per AA (since each AA handles fewer number of attributes per key issuing). In addition, the multiauthority ABE is resilient to compromise of up to N 1 2 AAs in a PUD,which solves the key-escrow problem. Furthermore, in ourframework user's role verification is much easier. Differentorganizations can form their own (sub)domains and becomeAAs to manage and certify different sets of attributes, whichis similar to divide and rule.

Using MA-ABE in the Public Domain

For the PUDs, our framework delegates the key management functions to multiple attribute authorities. In order to achieve stronger privacy guarantee for data owners, the Chase-Chow (CC) MA-ABE scheme [21] is used, where each authority governs a disjoint set of attributes distributively. It is natural to associate the ciphertext of a PHR document with an owner-specified access policy for users from PUD.

However, one technical challenge is that CC MA-ABE is essentially a KP-ABE scheme, where the access policies areenforced in users' secret keys, and those key-policies do notdirectly translate to document access policies from the owners' points of view. By our design, we show that by agreeing upon the formats of the key-policies and the rules of specifying which attributes are required in the ciphertext, the CC MA-ABE can actually support owner-specified document access policies with some degree of flexibility

Setup. In particular, the AAs first generate the MKs and PK using setup as in CC MA-ABE. The kth AA defines a disjoint set of role attributes UUk, Table-1,which are relatively static properties of the public users. These attributes are classified by their types, such as profession and license status, medical specialty, and affiliation where each type has multiple possible values. Basically, each AA monitors a disjoint subset of attribute types. For example, in the healthcare domain, the AMA may issue medical professional licenses like "physician,""M.D.," "nurse," "entry-level license," etc., the ABMS could certify specialties like "internal medicine," "surgery,"etc; and AHA may define user affiliations such as "hospitalA" and "pharmacy D." In order to represent the "do not care" option for the owners, we add one wildcard attribute in each type of the attributes.

TABLE 1
Frequently Used Notations

$\mathcal{U}_D, \mathcal{U}_R$	The attribute universes for data and roles
$\mathcal{T}, L(\mathcal{T})$	A user access tree and its leaf node set
A_k^C	Attributes in the ciphertext (from the kth AA)
A_k^u	User u's attributes given by the kth AA
A, a	An attribute type, a specific attribute value of that type
$\mathcal{P}$	Access policy for a PHR document
P	A key-policy assigned to a user
MK, PK	Master key and public key in ABE
SK	A user's secret key in ABE
$rk_j^{(k)}$	Proxy re-key for attribute j and version k

This primary-type based attribute association is illustratedin Fig. 2. Note that there is a "horizontal association"between two attributes belonging to different types assigned to each user. For example, in the first AA (AMA) "license status" is associated with "profession," and "profession" is a primary type. That means, a physician's possible set of license status do not intersect with that of a nurse's, or a pharmacist's. An "M.D." license is always associated with "physician," while "elderly's nursing licence" is always associated with "nurse." Thus, if thesecond level key policy within the AMA is "1 out of n1 ^ 1 out of n2," a physician would receive a key like "(physician OR *) AND (M.D. OR *)" (recall the assumption that each user can only hold at most one role attribute in each type),nurse's will be like "(nurse OR *) AND (elderly's nursing licence OR *)." Meanwhile, the encryptor can be made aware of this correlation, so she may include the attribute set: {physician, M.D., nurse, elderly's nursing licence} during encryption.

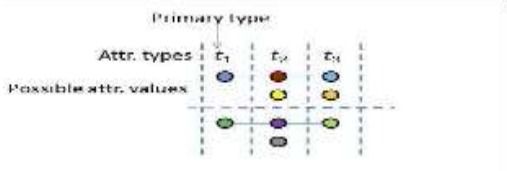

Fig 3: Enhanced Key policy generation rule

Due to the attribute correlation, the set of users that can have access to this file can only possess one out of two sets of possible roles, which means the following policy is enforced: "(physician AND M.D.) OR (nurse AND elderly's nursing licence)." The direct consequence is it enables a disjunctive normal form (DNF) encryptor access policy to appear at the second level. If the encryptor wants to enforce such a DNF policy under an AA, she can simply include all the attributes in that policy in the ciphertext. Furthermore, if one wants to encrypt with wildcard attributes in the policy, say: "(physician AND M.D.) OR (nurse ANDany nursing license)" the same idea can be used,i.e., we can simply correlate each "profession" attribute with its proprietary "*" attribute.

In this above, we present a method to enforce owner's access policy during encryption, which utilizes the MAABE scheme in a way like CP-ABE. The essential idea is todefine a set of key-generation rules and encryption rules.There are two layers in the encryptor's access policy, the first one is across different attribute authorities while thesecond is across different attributes governed by the same AA. For the first layer, conjunctive policy is enabled; for the second, either k-out-of-n or DNF policy are supported. We exploit the correlations among attribute types under an AA to enable the extended second-level DNF policy.

7.SECURITY ANALYSIS

The results are shown in Table 3. It can be seen that, our scheme achieves high privacy guarantee and on-demand revocation. The conjunctive policy restriction only applies forPUD, while in PSD a user's access structure can still bearbitrary monotonic formula. In comparison with the RNSscheme, in RNS the AAs are independent with each other,while in our scheme the AAs issue user secret keyscollectively and interactively. Also, theRNSscheme supportsarbitrary monotonic Boolean formula as file access policy.However, our user revocation method is more efficient interms of communication overhead. In RNS, upon eachrevocation event, the data owner needs to recompute andsend new ciphertext components corresponding to revokedattributes to all the remaining users. In our scheme, such interaction is not needed. In addition, our proposed framework specifically addresses the access requirements in cloudbased health record management systems by logically dividing the system into PUD and PSDs, which considersboth personal and professional PHR users. Our revocation methods for ABE in both types of domains a

8. CONCLUSION

In this paper, we have proposed a novel framework of secure sharing of personal health records in cloud computing.Considering partially trustworthy cloud servers, we argue that to fully realize the patient-centric concept, patients shall have complete control of their own privacy through encrypting their PHR files to allow fine-grained access. The framework addresses the unique challenges brought by multiple PHR owners and users, in that we greatly reduce the complexity of key management whileenhance the privacy guarantees compared with previousworks. We utilize ABE to encrypt the PHR data, so that patients can allow access not only by personal users, but also various users from public domains with different professional roles, qualifications, and affiliations. Furthermore, we enhance an existing MA-ABE scheme to handle efficient and on-demand user revocation, and prove its security. Through implementation and simulation, we showthat our solution is both scalable and efficient.

9. REFERENCES

[1] M. Li, S. Yu, K. Ren, and W. Lou, "Securing Personal Health Records in Cloud Computing: Patient-Centric and Fine-Grainedn Data Access Control in Multi-Owner Settings," Proc. Sixth Int'l ICST Conf. Security and Privacy in Comm. Networks (SecureComm '10),pp. 89-106, Sept. 2010.

[2]H. Lo¨ hr, A.-R. Sadeghi, and M. Winandy, "Securing the E-HealthCloud," Proc. First ACM Int'l Health Informatics Symp. (IHI '10),pp. 220-229, 2010.

[3] M. Li, S. Yu, N. Cao, and W. Lou, "Authorized Private KeywordSearch over Encrypted Personal Health Records in CloudComputing," Proc. 31st Int'l Conf. Distributed Computing Systems(ICDCS '11), June 2011.

[4] "The Health Insurance Portability and Accountability Act,"http://www.cms.hhs.gov/HIPAAGenInfo/01_Overview.asp,2012.

[5] "Google, Microsoft Say Hipaa Stimulus Rule Doesn't Apply toThem," http://www.ihealthbeat.org/Articles/2009/4/8/, 2012.

[6] "At Risk of Exposure - in the Push for Electronic Medical Records,Concern Is Growing About How Well Privacy Can Be Safeguarded,"http://articles.latimes.com/2006/jun/26/health/he-privacy26, 2006.

[7] K.D. Mandl, P. Szolovits, and I.S. Kohane, "Public Standards andPatients' Control: How to Keep Electronic Medical RecordsAccessible but Private," BMJ, vol. 322, no. 7281, pp. 283-287,Feb. 2001.

[8] J. Benaloh, M. Chase, E. Horvitz, and K. Lauter, "PatientControlled Encryption: Ensuring Privacy of Electronic

MedicalRecords," Proc. ACM Workshop Cloud Computing Security(CCSW '09), pp. 103-114, 2009.

[9] S. Yu, C. Wang, K. Ren, and W. Lou, "Achieving Secure, Scalable,and Fine-Grained Data Access Control in Cloud Computing,"Proc. IEEE INFOCOM '10, 2010.

[10] C. Dong, G. Russello, and N. Dulay, "Shared and SearchableEncrypted Data for Untrusted Servers," J. Computer Security,vol. 19, pp. 367-397, 2010.

[11] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-BasedEncryption for Fine-Grained Access Control of Encrypted Data,"Proc. 13th ACM Conf. Computer and Comm. Security (CCS '06),pp. 89-98, 2006.

[12] M. Li, W. Lou, and K. Ren, "Data Security and Privacy in WirelessBody Area Networks," IEEE Wireless Comm. Magazine, vol. 17,no. 1, pp. 51-58, Feb. 2010.

[13] A. Boldyreva, V. Goyal, and V. Kumar, "Identity-Based Encryptionwith Efficient Revocation," Proc. 15th ACM Conf. Computer and Comm. Security (CCS), pp. 417-426, 2008.

[14] L. Ibraimi, M. Petkovic, S. Nikova, P. Hartel, and W. Jonker,"Ciphertext-Policy Attribute-Based Threshold Decryption withFlexible Delegation and Revocation of User Attributes," 2009.

[15] S. Yu, C. Wang, K. Ren, and W. Lou, "Attribute Based DataSharing with Attribute Revocation," Proc. Fifth ACM Symp.

International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 63 - 69, 2015, ISSN:- 2319–8656

On pgrα Homeomorphisms in Intuitionistic Fuzzy Topological Spaces

M. Thirumalaiswamy
Department of Mathematics,
NGM college, Pollachi – 642 001,
Tamil Nadu, India.

A. Surya
Department of Mathematics,
PA Polytechnic College, Pollachi- 642 002,
Tamil Nadu, India.

Abstract: In this paper, we introduce intuitionistic fuzzy pgrα closed mapping, intuitionistic fuzzy pgrα open mapping, intuitionistic fuzzy pgrα homeomorphisms and study some of their properties in intuitionistic fuzzy topological spaces.

Keywords and Phrases: Intuitionistic fuzzy topology, intuitionistic fuzzy set, intuitionistic fuzzy pgrα continuous, intuitionistic fuzzy pgrα open mapping, intuitionistic fuzzy pgrα closed mapping and intuitionistic fuzzy pgrα homeomorphisms.

2010 Mathematics Subject Classification: 54A40, 03F55

1.INTRODUCTION

The concept of fuzzy sets was introduced by Zadeh [16] and later Atanassov [1] generalized this idea to intuitionistic fuzzy sets by using the notation of fuzzy sets. On the other hand Coker [2] introduced intuitionistic fuzzy topological spaces using the notion of intuitionistic fuzzy sets. In this paper, we introduce intuitionistic fuzzy pgrα-closed mapping, intuitionistic fuzzy pgrα open mapping, intuitionistic fuzzy pgrα homeomorphisms and study some of their properties.

2. PRELIMINARIES

Throughout this paper, (X,τ), (Y,σ) and (Z,γ) (or simply X, Y and Z) denote the intuitionistic fuzzy topological spaces (IFTS for short) on which no separation axioms are assumed unless otherwise explicitly mentioned. For a subset A of X, the closure, the interior and the complement of A are denoted by cl(A), int(A) and A^c respectively. We recall some basic definitions that are used in the sequal.

2.1. Definition [1]

Let X be a nonempty set. An intuitionistic fuzzy set (IFS for short) A in X is an object having the form A= $\langle x, \mu_A, \nu_A; x \epsilon X\rangle$ where the functions μ_A: X→ [0,1] and ν_A : X → [0,1] denote the degree of membership (namely $\mu_A(x)$) and the degree of non membership (namely $\nu_A(x)$) of each element $x \in X$ to the set A, respectively, and $0 \leq \mu_A(x) + \nu_A(x) \leq 1$ for each $x \in X$. Denote by IFS(X), the set of all intuitionistic fuzzy sets in X.

2.2. Definition [1]

Let A and B be IFSs of the form A = $\{\langle x, \mu_A(x), \nu_A(x)\rangle : x \in X\}$ and B= $\{\langle x, \mu_B(x), \nu_B(x)\rangle : x \in X\}$. Then

1. A⊆B if and only if $\mu_A(x) \leq \mu_B(x)$ and $\nu_A(x) \geq \nu_B(x)$ for all $x \in X$,
2. A=B if and only if A⊆B and B⊆ A,
3. A^c=$\{\langle x, \nu_A(x), \mu_A(x)\rangle : x \in X\}$,
4. A∩B=$\{\langle x, \mu_A(x) \wedge \mu_B(x), \nu_A(x) \vee \nu_B(x)\rangle : x \in X\}$,
5. A∪B=$\{\langle x, \mu_A(x) \vee \mu_B(x), \nu_A(x) \wedge \nu_B(x)\rangle : x \in X\}$.

For the sake of simplicity, we shall use the notation A= $\langle x, \mu_{A,} \nu_A\rangle$ instead of A= $\{\langle x, \mu_A(x), \nu_A(x)\rangle; x \in X\}$. The intuitionistic fuzzy sets $0_\sim$= $\{\langle x, 0,1\rangle : x \in X\}$ and $1_\sim$= $\{\langle x, 1,0\rangle : x \in X\}$ are respectively the empty set and the whole set of X.

2.3. Definition [2]

An intuitionistic fuzzy topology (IFT for short) on X is a family τ of IFSs in X satisfying the following axioms.

1. $0_\sim, 1_\sim \epsilon\ \tau$,
2. $G_1 \cap G_2 \in \tau$ for any $G_1, G_2 \epsilon\ \tau$,
3. $\cup G_i \in \tau$ for any family $\{G_i : i \in J\} \subseteq \tau$.

In this case the pair (X,τ) is called an intuitionistic fuzzy topological space (IFTS for short) and any IFS in τ is known as an intuitionistic fuzzy open set (IFOS for short) in X. The complement A^c of an IFOS A in an IFTS (X,τ) is called an intuitionistic fuzzy closed set (IFCS for short) in X.

2.4. Definition [2]

Let (X,τ) be an IFTS and A=$\langle x, \mu_{A,} \nu_A\rangle$ be an IFS in X. Then

1. int(A) = ∪ $\{G : G$ is an IFOS in X and G ⊆ $A\}$,
2. cl(A) = ∩ {K: K is an IFCS in X and A ⊆ $K\}$.

For any IFS A in (X,τ), we have cl (A^c) = $(\text{int}(A))^c$ and $\text{int}(A^c) = (\text{cl}(A))^c$

2.5. Definition [3]

An IFS A=$\langle x, \mu_{A,} \nu_A\rangle$in an IFTS (X,$\tau$) is said to be an

1. intuitionistic fuzzy semi closed set (IFSCS for short) if int(cl(A)) ⊆ A,
2. intuitionistic fuzzy semi open set (IFSOS for short) if $A \subseteq$ cl(int((A))
3. intuitionistic fuzzy pre closed set (IFPCS for short) if cl(int((A)) ⊆ A,

4. intuitionistic fuzzy pre open set (IFPOS for short) if $A \subseteq \text{int}(\text{cl}(A))$,
5. intuitionistic fuzzy regular closed set (IFRCS for short) if cl(int(A))= A,
6. intuitionistic fuzzy regular open set (IFROS for short) if A= int(cl(A)),
7. intuitionistic fuzzy α closed set (IFαCS for short) if cl(int(cl(A))) $\subseteq A$,
8. intuitionistic fuzzy α open set (IFαOS for short) if A $\subseteq$ int(cl(int(A))).

2.6. Definition

Let A=$\langle x, \mu_{A,} v_A \rangle$ be an IFS in an IFTS (X,τ). Then
1. αint(A) = $\cup$ {G: G is an IFαOS in X and G$\subseteq$ A} and αcl(A) = $\cap$ {K: K is an IFαCS in X and A$\subseteq$ K}**[9]**,
2. $p$int(A = $\cup$ {G: G is an IFPOS in X and G$\subseteq$ A} and $p$cl(A) = $\cap$ {K: K is an IFPCS in X and A$\subseteq$ K}**[3]**,
3. sint(A) = $\cup$ {G: G is an IFSOS in X and G$\subseteq$ A} and $s$cl(A) = $\cap$ {K: K is an IFSCS in X and A$\subseteq$ K}**[9]**.

2.7. Definition [15]

An IFS A of an IFTS (X,τ) is called an intuitionistic fuzzy regular α open set ((IFRαOS for short) if there exist an IFROS U such that U $\subseteq$ A $\subseteq$ αcl(U).

2.8. Definition

An IFS A=$\langle x, \mu_{A,} v_A \rangle$ in an IFTS (X,τ) is called an

1. intuitionistic fuzzy generalized closed set (IFGCS for short) if cl(A) $\subseteq$ U , whenever A$\subseteq$ U and U is an IFOS in X **[13]**,
2. intuitionistic fuzzy α generalized closed set (IFαGCS for short) if αcl(A) $\subseteq$ U , whenever A$\subseteq$ U and U is an IFOS in X **[9]**,
3. intuitionistic fuzzy weakly generalized closed set (IFWGCS for short) if cl(int(A)) $\subseteq$ U, whenever A$\subseteq$ U and U is an IFOS in X **[7]**,
4. intuitionistic fuzzy generalized semi closed set (IFGSCS for short) if scl(A) $\subseteq$ U ,whenever A$\subseteq$ U and U is an IFOS in X **[11]**,
5. intuitionistic fuzzy regular generalized α closed set (IFRGαCS for short) if αcl(A) $\subseteq$ U, whenever A$\subseteq$ U and U is an IFRαOS in X **[5]**.

An IFS A is said to be an intuitionistic fuzzy generalized open set (IFGOS for short), intuitionistic fuzzy α generalized open set (IFαGOS for short), intuitionistic fuzzy weakly generalized open set (IFWGOS for short), intuitionistic fuzzy generalized semi open set (IFGSOS for short) and intuitionistic fuzzy regular generalized α open set (IFRGαOS for short) if the complement of A is an IFGCS, IFαGCS, IFWGCS, IFGSCS, and IFRGαCS respectively

2.9. Definition [6]

An IFS A = $\langle x, \mu_{A,} v_A \rangle$ in an IFTS (X,τ) is called an intuitionistic fuzzy regular weakly generalized closed set (IFRWGCS for short) if cl(int(A)) $\subseteq$ U, whenever A $\subseteq$ U and U is an IFROS in X. An IFS A is called an intuitionistic fuzzy regular weakly generalized open set (IFRWGOS for short) in X if the complement of A is an IFRWGCS in X.

2.10. Definition [14]

An IFS A in an IFTS (X,τ) is said to be an intuitionistic fuzzy pgrα closed set (IF pgrα CS for short) if pcl (A) $\subseteq$ U whenever A $\subseteq$ U and U is an IFRαOS in X. The family of all IF pgrαCSs of an IFTS (X, τ) is denote by IFpgrαC(X).

2.11. Definition [3]

A mapping f: (X,τ) $\rightarrow$ (Y,σ) is called an intuitionistic fuzzy continuous mapping (IF continuous mapping for short) if f^{-1}(B) is an IFOS in (X,τ) for every IFOS B of (Y,σ).

2.12. Definition [4]

Let f: (X,τ) $\rightarrow$ (Y,σ) be a mapping. Then f is said to be an
1. intuitionistic fuzzy α continuous mapping (IF α continuous mapping for short) if f^{-1}(B) ϵ IFαO(X) for every B$\epsilon\sigma$,
2. intuitionistic fuzzy pre continuous mapping (IFP continuous mapping for short) if f^{-1}(B) ϵ IFPO(X) for every B ϵ σ.

2.13. Defintion [5]

Let f: (X,τ) $\rightarrow$(Y,σ) be a mapping. Then f is said to be an intuitionistic fuzzy regular generalized α continuous mapping (IFRGα continuous mapping for short) if f^{-1}(B) is an IFRGαCS in (X,τ) for every IFCS B of (Y,σ).

2.14. Definition [8]

Let f: (X,τ) $\rightarrow$(Y,σ) be a mapping. Then f is said to be an intuitionistic fuzzy regular weakly generalized continuous mapping (IFRWG continuous mapping for short) if f^{-1}(B) is an IFRWGCS in (X,τ) for every IFCS B of (Y,σ).

2.15. Definition [14]

A mapping f: (X,τ) $\rightarrow$(Y,σ) is called an intuitionistic fuzzy pgrα continuous (IFpgrα continuous for short) mapping if f^{-1}(V) is an IFpgrαCS in (X,τ) for every IFCS V of (Y,σ).

2.16. Definition [12]

A mapping f: (X,τ) $\rightarrow$(Y,σ) is called an intuitionistic fuzzy closed mapping (IFCM for short) if f(A) is an IFCS in Y for each IFCS A in X.

2.17. Definition [12]

A mapping f: (X,τ) $\rightarrow$(Y,σ) is called an
1. intuitionistic fuzzy semi open mapping (IFSOM for short) if f(A) is an IFSOS in Y for each IFOS A in X.
2. intuitionistic fuzzy pre open mapping (IFPOM for short) if f(A) is an IFPOS in Y for each IFOS A in X.

2.18. Definition [5]

A mapping f: (X,τ) $\rightarrow$(Y,σ) is an intuitionistic fuzzy rgα closed mapping (IFRGα CM for short) if image of every IFCS of X is an IFRGα CS in Y.

2.19. Definition [2]

Let f: $(X,\tau) \to (Y,\sigma)$ be a mapping. Then f is said to be intuitionistic fuzzy homeomorphism (IF homeomorphism for short) if f and f^{-1} are IF continuous mappings.

2.20. Definition [10]

Let f: $(X,\tau) \to (Y,\sigma)$ be a mapping. Then f is said to be intuitionistic fuzzy α homeomorphism (IF α homeomorphism for short) if f and f^{-1} are IF α continuous mappings.

3. INTUITIONISTIC FUZZY pgrα CLOSED MAPPING AND INTUITIONISTIC FUZZY pgrα OPEN MAPPING

In this section we introduce intuitionistic fuzzy pgrα closed mapping, intuitionistic fuzzy pgr α open mapping and investigate some of its properties.

3.1. Definition

A mapping f: $(X,\tau) \to (Y,\sigma)$ is called an intuitionistic fuzzy pgrα closed mapping (IF pgrα CM for short) if f(A) is an IF pgrα CS in Y for each IFCS A in X.

3.2. Example

Let X={a,b}, Y={u,v} and G_1=<x,(0.8,0.7),(0.2,0.2)>, G_2=<x,(0.3,0.3),(0.7,0.7)>. Then τ= {$0_{\sim}$, G_1, $1_{\sim}$} and σ ={$0_{\sim}$,G_2, $1_{\sim}$} are IFTs on X and Y respectively. Define a mapping f: (X, τ) → (Y, σ) by f(a)=u and f(b)=v. Then f is an IF pgrα CM.

3.3. Definition

A mapping f: $(X,\tau) \to (Y,\sigma)$ is called an intuitionistic fuzzy pgrα open mapping (IF pgrα OM for short) if f(A) is an IF pgrα OS in Y for each IFOS A in X.

3.4. Definition

A mapping f: $(X,\tau) \to (Y,\sigma)$ is called an intuitionistic fuzzy ipgrα closed mapping (IFipgrα CM for short) if f(A) is an IF pgrα CS in Y for each IF pgrα CS A in X.

3.5. Example

Let X={a,b}, Y={u,v} and G_1=<x,(0.3,0.3),(0.7,0.7)>, G_2=<x,(0.8,0.7),(0.2,0.2)>. Then τ= {$0_{\sim}$, G_1, $1_{\sim}$} and σ ={$0_{\sim}$,G_2, $1_{\sim}$} are IFTs on X and Y respectively. Define a mapping f: (X, τ) → (Y, σ) by f(a)=u and f(b)=v. Then f is an IFipgrα CM.

3.6. Definition

A mapping f: $(X,\tau) \to (Y,\sigma)$ is called an intuitionistic fuzzy ipgrα open mapping (IFipgrα OM for short) if f(A) is an IF pgrα OS in Y for each IF pgrα OS A in X.

3.7. Definition

Let (X,τ) be an IFTS and A= $\langle x, \mu_A, \nu_A \rangle$ be an IFS in X. Then pgrα -interior of A (pgrαint(A) for short) and pgrα -closure of A (pgrαcl(A) for short) are defined as

1. pgrαint(A) = $\cup$ {$G: G$ is an IF pgrα OS in X and G $\subseteq A$},

2. pgrαcl(A) = $\cap$ {K: K is an IF pgrα CS in X and A $\subseteq K$}.

3.8. Theorem

Every IFCM is an IF pgrα CM but not conversely.

Proof:

Let f: (X, τ) → (Y, σ) be an IF closed mapping. Let A be an IFCS in X. Then f(A) is IFCS in Y. This implies that f(A) is an IF pgrα CS inY. Hence f is an IF pgrα closed mapping.

3.9. Example

In Example 3.2, f: (X, τ) → (Y, σ) is an IF pgrα CM but not an IFCM.

3.10. Theorem

Every IF α CM is an IF pgrα CM but not conversely.

Proof:

Let f: (X, τ) → (Y, σ) be an IFα closed mapping. Let A be an IFCS in X. Then f(A) is IF α CS in Y. This implies that f(A) is an IF pgrα CS in Y. Hence f is an IF pgrα closed mapping.

3.11. Example

In Example 3.2, f: (X, τ) → (Y, σ) is an IF pgrα CM but not an IF α CM.

3.12. Theorem

Every IFPCM is an IF pgrα CM but not conversely.

Proof:

Let f: (X, τ) → (Y, σ) be an IFP closed mapping. Let A be an IFCS in X. Then f(A) is IFPCS in Y. This implies that f(A) is an IF pgrα CS in Y. Hence f is an IF pgrα closed mapping.

3.13. Example

Let X={a,b}, Y={u,v} and G_1=<x,(0.3,0.1),(0.7,0.9)>, G_2=<x,(0.7,0.9),(0.3,0.1)>. Then τ= {$0_{\sim}$, G_1, $1_{\sim}$} and σ ={$0_{\sim}$, G_2, $1_{\sim}$} are IFTs on X and Y respectively. Define a mapping f: (X, τ) → (Y, σ) by f(a)=u and f(b)=v. Then f is an IF pgrα CM but not IFPCM.

3.14. Theorem

Every IFRG α CM is an IF pgrα CM but not conversely.

Proof:

Let f: (X, τ) → (Y, σ) be an IFRG α CM. Let A be an IFCS in X. Then f(A) is IFRG α CS in Y. This implies that f(A) is an IF pgr α CS in Y. Hence f is an IF pgr α closed mapping.

3.15. Example
In Example 3.2, f: (X, τ) → (Y, σ) is an IF pgrα CM but not an IFRG α CM.

3.16. Theorem
Every IF ipgrα CM is an IF pgrα CM but not conversely.

Proof:
Let f: (X, τ) → (Y, σ) be an IF ipgrα CM. Let A be an IFCS in X. Then A is IF pgrα CS in Y. This implies that f(A) is an IF pgrα CS in Y. Hence f is an IF pgrα closed mapping.

3.17. Example
Let X={a,b}, Y={u,v} and G_1=<x,(0.8,0.7),(0.2,0.2)>, G_2=<x,(0.3,0.3),(0.7,0.7)>. Then τ= {0$_{\sim}$, G_1, 1$_{\sim}$} and σ ={0$_{\sim}$, G_2, 1$_{\sim}$} are IFTs on X and Y respectively. Define a mapping f: (X, τ) → (Y, σ) by f(a)=u and f(b)=v. Then f is an IF pgrα CM but not IF ipgrα CM.

3.18. Theorem
A mapping f: (X, τ) → (Y, σ) is IF pgrα closed mapping if and only if for each subset S of Y and for each IFOS U containing f^{-1}(S) there is an IF pgrα OS V of Y such that S⊆V and f^{-1}(V) ⊆ U.

Proof:
Suppose f is an IF pgrα closed. Let S be a subset of Y and U is an IFOS of X such that f^{-1}(S) ⊆ U. Then V = Y-f(X-U) is an IF pgr α OS containing S such that f^{-1}(V) ⊆ U.

Conversely, suppose that F is an IFCS in X. Then f^{-1}(Y-f(F)) ⊆ X-F, X-F is an IFOS in X. By hypothesis, there is an IF pgrα OS V of Y such that Y-f(F) ⊆ V and f^{-1}(V) ⊆ X-F.Therefore F ⊆ X-f^{-1}(V). Hence Y-V⊆ f(F) ⊆ f(X-f^{-1}(V)) ⊆ Y-V, which impiles f(F)=Y-V. Since Y-V is an IF pgrα CS in Y, f(F) is an IF pgrα CS in Y and therefore f is an IF pgrα closed mapping.

3.19. Theorem
If f: (X, τ) → (Y, σ) is an IF pgrα CM and A is an IFCS of X, then f_A:A→Y is IF pgrα CM.

Proof:
Let B⊆A be ab IFCS in A, then B is an IFCS in X. Since A is an IFCS in X, f(B) is an IF pgrα CS in Y as f is IF pgrα CM. But f(B)= f_A(B). So f_A(B) is an IF pgrα CS in Y. Therefore f_A is an IF pgrα CM.

3.20. Remark

Composition of two IF pgrα CMs need not be an IF pgrα CM.

3.21. Example
Let X={a,b}, Y={c,d} and Z={u,v}. Let G_1=<x,(0.5,0.6),(0.5,0.4)>, G_2=<x,(0.6,0.1),(0.4,0.3)> and G_3=<x,(0.4,0.4),(0.6,0.6)>,. Then τ= {0$_{\sim}$, G_1, 1$_{\sim}$}, σ ={0$_{\sim}$, G_2, 1$_{\sim}$} and γ ={0$_{\sim}$, G_3, 1$_{\sim}$} are IFTs on X and Y respectively. Define a mapping f: (X, τ) → (Y, σ) by f(a)=c and f(b)=d and g: (Y, σ) → (Z,γ) by g(c)=u and g(d)=v. Then f and g are an IF pgrα CM. But their composition gof: (X, τ) → (Z,γ) need not be an IF pgrα CM.

3.22. Theorem
If f: (X, τ) → (Y, σ) is an IFCM and g: (Y, σ) → (Z,γ) is an IF pgrα CM, then gof: (X, τ) → (Z,γ) is an IF pgrα CM.

Proof:
Let H be an IFCS in X. Then f(H) is an IFCS. But (gof)(H)=g(f(H)) is an IF pgrα CS as g ia an IF pgrα CM. Thus gof is an IF pgrα CM.

3.23. Theorem
If f: (X, τ) → (Y, σ) is a bijective mapping, then the following statements are equivalent

1. f is an IF pgrα OM.

2. f ia an IF pgrα CM.

3. f^{-1}: (Y, σ) →(X, τ) is an IF pgrα continuous.

Proof:
(1)⇒(2) Let U be an IFCS in X and f be an IF pgrα OM. Then X-U is an IFOS in X. By hypothesis, we get f(X-U) is an IF pgrα OS in Y. That is Y-f(X-U)-f(U) is an IF pgrα in Y.

(2) ⇒(3) L is an et U ne an IFCS in X. By assumption, f(U) IF pgrα CS in Y. As f(U)=$(f^{-1})^{-1}$(U), f^{-1} is an IF pgrα continuous.

(3) ⇒ (1) Let U be an IFOS in X. By assumption $(f^{-1})^{-1}$(U) =f(U). That is f(U) is an IF pgrα OS in Y. Hence f is an IF pgrα OM.

3.24. Definition
A space (X, τ) is called an IFpgr$\alpha T_{1/2}$ space if every IF pgrα CS is an IFα CS.

3.25. Definition
A space (X, τ) is called an IFpgr$T_{1/2}$ space if every IF pgrα CS is an IFCS.

3.26. Definition
A mapping f: (X, τ) → (Y, σ) is called an intuitionistic fuzzy pgrα irresolute (IF pgrα irresolute in short) mappings if f^{-1}(V) is an IF pgrα CS in (X, τ) for every IF pgrα CS V of (Y, σ).

3.27. Theorem
For any bijective mapping f: (X, τ) → (Y, σ), then the following are equivalent

1. f^{-1}: (Y, σ) →(X, τ) is an IF pgrα irresolute mapping
2. f is an IF ipgrα OM
3. f is an IF ipgrα CM

Proof:

(1)⇒(2) Let U be an IF pgrα OS in X. By (1), $(f^{-1})^{-1}$(U) =f(U) is an IF pgrα OS in Y So f is an IF ipgrα OM.

(2) ⇒(3) Let V be an IF pgrα CS in X. By (2), f(X-V)= Y-f(V) is an IF pgrα OS in Y. That is f(V) is an IF pgrα CS in Y and so f is an IF ipgrα CM.

(3) ⇒ (1) Let V be an IF pgr α CS in X. By (3), f(V)= $(f^{-1})^{-1}$(V) is an IF pgrα CS in Y. Hence (1) holds.

3.28. Theorem

If f: (X, τ) → (Y, σ) and g: (Y, σ) → (Z, γ) are IF ipgrα CM, then gof: (X, τ) → (Z,γ) is an IF ipgrα CM.

Proof:

Let V be an IF pgrα CS in X. Since f is an IF ipgrα CM, f(V) is an IF pgrα CS in Y. Then g(f(V)) is an IF pgrα CS in Z. Hence gof is an IF ipgrα CM.

3.29. Theorem

If f: (X, τ) → (Y, σ) is an IF pgrα CM and g: (Y, σ) → (Z,γ) is an IF ipgrα CM, then gof: (X, τ) → (Z,γ) is an IF pgrα CM.

Proof:

Let V be an IFCS in X. Since f is an IF pgrα CM, f(V) is an IF pgrα CS in Y. Then g(f(V)) is an IF pgrα CS in Z. Hence gof is an IF pgrα CM.

4 INTUITIONISTIC FUZZY pgrα HOMEOMORPHISM

In this section, we introduce the concept of intuitoinstic fuzzy pgrα homeomorphism, intuitionistic fuzzy ipgrα homeomorphism and study some of their properties.

4.1. Definition

A bijective mapping f: (X, τ) → (Y, σ) is called an intuitionistic fuzzy pgr α homeomorphism (IF pgr α homeomorphism in short) if f and f^{-1} are IF pgrα continuous mapping.

4.2. Example

Let X={a,b}, Y={u,v} and G_1=<x,(0.3,0.2),(0.6,0.7)>, G_2=<x,(0.8,0.9),(0.2,0.1)>. Then τ= {0$_\sim$, G_1, 1$_\sim$} and σ ={0$_\sim$, G_2, 1$_\sim$} are IFTs on X and Y respectively. Define a mapping f: (X, τ) → (Y, σ) by f(a)=u and f(b)=v. Then f is an IF pgrα continuous mapping and f^{-1} is also an IF pgrα continuous mapping. Therefore f is an IF pgr α homeomorphism.

4.3. Theorem

Every IF homeomorphism is an IF pgrα homeomorphism but not conversely.

Proof:

Let f: (X, τ) → (Y, σ) be an IF homeomorphism. Then f and f^{-1} are IF continuous mapping. This implies that f and f^{-1} are IF pgrα continuous mapping, that is the mapping f is an IF pgrα homeomorphism.

4.4. Example

Let X={a,b}, Y={u,v} and G_1=<x,(0.2,0.3),(0.8,0.7)>, G_2=<x,(0.6,0.8),(0.3,0.2)>. Then τ= {0$_\sim$, G_1, 1$_\sim$} and σ ={0$_\sim$, G_2, 1$_\sim$} are IFTs on X and Y respectively. Define a mapping f: (X, τ) → (Y, σ) by f(a)=u and f(b)=v. Then f is an IF pgrα continuous mapping and f^{-1} is also an IF pgrα continuous mapping. Therefore f is an IF pgr α homeomorphism but not IF homeomorphism.

4.5. Theorem

Every IF α homeomorphism is an IF pgrα homeomorphism but not conversely.

Proof:

Let f: (X, τ) → (Y, σ) be an IF α homeomorphism. Then f and f^{-1} are IF α continuous mapping. This implies that f and f^{-1} are IF pgr α continuous mapping, that is the mapping f is an IF pgrα homeomorphism.

4.6. Example

Let X={a,b}, Y={u,v} and G_1=<x,(0.5,0.1),(0.5,0.9)>, G_2=<x,(0.2,0.2),(0.7,0.8)>. Then τ= {0$_\sim$, G_1, 1$_\sim$} and σ = {0$_\sim$, G_2, 1$_\sim$} are IFTs on X and Y respectively. Define a mapping f: (X, τ) → (Y, σ) by f(a)=u and f(b)=v. Therefore f is an IF pgrα homeomorphism. Consider the IFCS A= <x,(0.7,0.8),(0.2,0.2)> in Y. Then f^{-1} (A) = <x,(0.7,0.8),(0.2,0.2)> is not IF α CS in X. This implies that f is not an IF α continuous mapping. Hence f is not an IF α homeomorphism.

4.7. Theorem

Let f: (X, τ) → (Y, σ) be an IF pgrα homeomorphism, then f is an IF homeomorphism if X and Y are IFpgr$T_{1/2}$ space.

Proof:

Let B be an IFCS in Y. Then f^{-1} (B) is an IF pgrα CS in X, by hypothesis. Since X is an IFpgr$T_{1/2}$ space, f^{-1} (B) is an IFCS in X. Hence f is an IF continuous mapping. By hypothesis f^{-1}: (Y, σ) →(X, τ) is a IF pgrα continuous mapping. Let A be an IFCS in X. Then $(f^{-1})^{-1}$(A) =f(A) is an IF pgrα CS in Y, by hypothesis. Since Y is an IFpgr$T_{1/2}$ space, f(A) is an IFCS in Y. Hence f^{-1} is an IF continuous mapping. Therefore the mapping f is an IF homeomorphism.

4.8. Theorem

Let f: (X, τ) → (Y, σ) be an IF pgrα homeomorphism, then f is an IF α homeomorphism if X and Y are IFpgr$\alpha T_{1/2}$ space.

Proof:

Let B be an IFCS in Y. Then f^{-1} (B) is an IF pgrα CS in X, by hypothesis. Since X is an IFpgr$\alpha T_{1/2}$ space, f^{-1} (B)

is an IF α CS in X. Hence f is an IF α continuous mapping. By hypothesis f^{-1}: (Y, σ) →(X, τ) is a IF pgrα continuous mapping. Let A be an IFCS in X. Then $(f^{-1})^{-1}$(A) =f(A) is an IF pgrα CS in Y, by hypothesis. Since Y is an IFpgr$\alpha T_{1/2}$ space, f(A) is an IF α CS in Y. Hence f^{-1} is an IF α continuous mapping. Therefore the mapping f is an IF α homeomorphism.

4.9. Theorem

Let f: (X, τ) → (Y, σ) be a bijective mapping. If f is an IF pgr α continuous mapping, then the following are equivalent

1.f is an IF pgrα CM.

2.f is an IF pgrα OM.

3.f is an IF pgrα homeomorphism**.**

Proof:

(1)⇒(2) Let f: (X, τ) → (Y, σ) be a bijective mapping and IF pgrα CM. This implies that f^{-1}: (Y, σ) →(X, τ) is an IF pgrα continuous mapping. That is, every IFOS in X is an IF pgrα OS in Y. Hence f^{-1} is an IF pgrα OM.

(2)⇒(3) Let f: (X, τ) → (Y, σ) be a bijective mapping and IF pgrα OM. This implies that f^{-1}: (Y, σ) →(X, τ) is an IF pgrα continuous mapping. Hence f and f^{-1} are IF pgrα continuous mapping. That is, f is an IF pgr α homeomorphism.

(3)⇒(1) Let f be an IF pgrα homeomorphism. That is, f and f^{-1} are IF pgrα continuous mappings. Since every IFCS in X is an IF pgrα CS in Y, f is an IF pgrα CM.

4.10. Definition

A bijective mapping f: (X, τ) → (Y, σ) is called an intuitionistic fuzzy ipgr α homeomorphism (IF ipgrα homeomorphism in short) if f and f^{-1} are IF pgrα irresolute mappings.

4.11. Theorem

Every IFipgr α homeomorphism is an IF pgrα homeomorphism but not conversely.

Proof:

Let f: (X, τ) → (Y, σ) be an IF ipgrα homeomorphism. Let B be IFCS in Y. This implies B is an IF pgrα CS in Y. By hypothesis f^{-1}(B) is an IF pgrα CS in X. Hence f is an IF pgrα continuous mapping. Similarly, we can prove f^{-1} is an IF pgrα continuous mapping. Hence f and f^{-1} are IF pgrα continuous mapping. This implies that the mapping f is an IF pgrα homeomorphism.

4.12. Example

Let X={a,b}, Y={u,v} and G_1=<x,(0.3,0.2),(0.7,0.8)>, G_2=<x,(0.9,0.9),(0.1,0.1)>. Then τ= {$0_{\sim}$, G_1, $1_{\sim}$} and σ ={$0_{\sim}$, G_2, $1_{\sim}$} are IFTs on X and Y respectively. Define a mapping f: (X, τ) → (Y, σ) by f(a)=u and f(b)=v. Therefore f is an IF pgr α homeomorphism. Consider the IFCS A=<x,(0.4,0.2),(0.6,0.8)> in Y. Clearly A is an IF pgrα CS. But f^{-1}(A) is not IF α CS in X. That is f is not an IF pgrα irresolute mapping. Hence f is not an IF ipgr α homeomorphism.

4.13. Theorem

The composition of two IF ipgrα homeomorphisms is IF ipgrα homeomorphism in general.

Proof:

Let f: (X, τ) → (Y, σ) and g: (Y, σ) → (Z,γ) be two any IF ipgrα homeomorphisms. Let A be an IF pgrα CS in Z. Then by hypothesis, g^{-1}(A) is an IF pgrα CS in Y. Again by hypothesis, f^{-1} (g^{-1}(A)) is an IF pgr α CS in X. Therefore gof is an IF pgrα irresolute mapping. Now let B be an IF pgrα CS in X. Then by hypothesis, f(B) is an IF pgrα CS in Y and also g(f(B)) is an IF pgrα CS in Z. This implies gof is an IF pgrα irresolute mapping. Hence gof is an IF ipgrα homeomorphism.

5. REFERENCES

[1] K.T. Atanassov, Intuitionistic fuzzy sets, Fuzzy Sets and Systems, 20(1986), 87-96

[2] D. Coker, An introduction to intuitionistic fuzzy topological spaces, Fuzzy Sets and systems, as (1997) 81-89.

[3] H. Gurcay, D. Coker and Es. A. Hayder, On fuzzy continuity in intuitionistic fuzzy topological fuzzy spaces, The Journal of Fuzzy Mathematics, 5(1997), 365-378

[4] A. Joung Kon Jeon, Young Bae Jun and Jin Han Park, Intuitionistic fuzzy alpha continuity and intuitionistic fuzzy pre continuity, International Journal of Mathematics and Mathematical Sciences, 19 (2005), 3091-31011.

[5] Jyoti Pandey Bajpal and S.S. Thakur, Intuitionistic fuzzy rg α continuity, Int. J. Contemp, Math. Sciences, 6(2011), 2335-2351.

[6] P. Rajarajeswari and L.Senthil Kumar, Regular weakly generalized closed sets in intuitionistic fuzzy topological spaces, International Journal of computer Applications, 43(18) (2012), 13-17.

[7] P. Rajarjeswari and R.Krishna Moorthy, On Intuitionistic fuzzy weakly generalized closed set and its applications, Int. J. Comput, Appl., 27(11)(2011), 9-13

[8] P.Rajrajeswari and L. Senthil Kumar, Regular weakly generalized continuous mappings in intuitionistic fuzzy topological spaces, International journal of Mathematical Archive, 3(5) (2012), 1957-1962.

[9] K. Sakthivel, Intuitionistic fuzzy alpha generalized continuous mapping and intuitionistic fuzzy alpha generalized irresolute mappings, Applied Mathematical Sciences, 4(2010), 1831-1842.

[10] K. Sakthivel, Alpha generalized homeomorphism in intuitionistic fuzzy topological spaces, Notes on Intuitionistic Fuzzy Sets, 17(2011), 30-36.

[11] R.Santhi and K. Sakthivel, Intuitionistic fuzzy generalized semi continuous mappings, Advances in Theoretical and Applied Mathematics, 59(2009), 73-82.

[12] Seok Jong Lee and Eun Pyo Lee, The Category of intuitionistic fuzzy topological spaces, Bull. Korean Math. Soc., 37(1) (2000), 63-76.

[13] S.S. Thakur and Rekha Chaturvedi, Generalized closed sets in intuitionistic fuzzy topology, The J.fuzzy Math., 16(3) (2008), 559-572.

[14] M. Thirumalaiswamy and A. Surya, On pgrα Closed Sets in Intuitionistic Fuzzy Topological Spaces,

[15] S. S. Thakur and Bajpai Pandey Jyoti, Intuitionistic Fuzzy r$g\alpha$ Closed Sets, International Journal of Fuzzy System and Rough System, 4(1) (2011),67-73.

[16] L.A. Zadeh, fuzzy sets, information and Control, 8(1955), 338-353.

International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 70 - 76, 2015, ISSN:- 2319–8656

Satellite Image Resolution Enhancement Technique Using DWT and IWT

E. Sagar Kumar
Dept of ECE (DECS),
Vardhaman College of Engineering,

MR. T. Ramakrishnaiah
Assistant Professor (Sr.Grade),
Vardhaman College of Engineering,

Abstract: Now a days satellite images are widely used In many applications such as astronomy and geographical information systems and geosciences studies .In this paper, We propose a new satellite image resolution enhancement technique which generates sharper high resolution image .Based on the high frequency sub-bands obtained from the dwt and iwt. We are not considering the LL sub-band here. In this resolution-enhancement technique using interpolated DWT and IWT high-frequency sub band images and the input low-resolution image. Inverse DWT (IDWT) has been applied to combine all these images to generate the final resolution-enhanced image. The proposed technique has been tested on satellite bench mark images. The quantitative (peak signal to noise ratio and mean square error) and visual results show the superiority of the proposed technique over the conventional method and standard image enhancement technique WZP.

Keywords: DWT; Interpolation;IWT; Resolution; WZP;

I.INTRODUCTION

Resolution has been frequently referred as an important aspect of an image. Images are being processed in order to obtain more enhanced resolution. One of the commonly used techniques for image resolution enhancement is Interpolation. Interpolation has been widely used in many image processing applications such as facial reconstruction, multiple description coding, and super resolution. There are three well known interpolation techniques, namely nearest neighbor interpolation, bilinear interpolation, and bi-cubic interpolation. The results are decomposed along the columns. This operation results in four decomposed sub band images referred to low(LL), low-high (LH), high-low (HL), and high-high (HH).The frequency components of those sub bands cover the full frequency spectrum of the original image. Image resolution enhancement using wavelets is a relatively new subject and recently many new algorithms have been proposed. Their estimation was carried out by investigating the evolution of wavelet transform extreme among the same type of sub bands. Edges identified by an edge detection algorithm in lower frequency sub bands were used to prepare a model for estimating edges in higher frequency sub bands and only the coefficients with significant values were estimated as the evolution of the wavelet coefficients. In many researches, hidden Markov has been also implemented in order to estimate the coefficients. The proposed technique has been compared with conventional and state-of-art image resolution enhancement techniques. The conventional techniques used are the following interpolation techniques: bilinear interpolation and bi-cubic interpolation, wavelet zero padding (WZP),DWT based super resolution (DWT SR)[1].According to the quantitative and qualitative experimental results, the proposed technique over performs the aforementioned conventional and state-of-art techniques for image resolution enhancement.

II. INTEGER WAVELET TRANSFORM (IWT)

Integer to integer wavelet transforms (IWT)[2] maps an integer data set into another integer data set. This transform is perfectly invertible and yield exactly the original data set. A one dimensional discrete wavelet transform is a repeated filter bank algorithm. The reconstruction involves a convolution with the syntheses filters and the results of these convolutions are added. In two dimensions, we first apply one step of the one dimensional transform to all rows. Then, we repeat the same for all columns. In the next step, we proceed with the coefficients that result from a convolution in both directions. As shown in figure 1, these steps result in four classes of coefficients: the (HH) coefficients represent diagonal features of the image, whereas (HL and LH) reflect vertical and horizontal information. At the coarsest level, we also keep low pass coefficients (LL). We can do the same decomposition on the LL quadrant up to log2(min (height, width)).Since the discrete wavelet transform allows independent processing of the resulting components without significant perceptible interaction To make the process of imperceptible embedding more effective. However, the used wavelet filters have floating point coefficients. Thus, when the input data consist of sequences of integers (as in the case for images), the resulting filtered outputs no longer consist of integers, which doesn't allow perfect reconstruction of the original image. However, with the introduction of Wavelet transforms that map integers to integers we are able to characterize the output completely with integers

.

III. CONVENTIONAL METHOD

As it was mentioned before, resolution is an important feature in satellite imaging, which makes the resolution enhancement of such images to be of vital importance as increasing the resolution of these images will directly affect the performance of the system using these images as input. The main loss of an image after being resolution enhanced by applying interpolation is on its high-frequency components, which is due to the smoothing caused by interpolation. Hence, in order to increase the quality of the enhanced image, preserving the edges is essential. In this paper, DWT has been employed in order to preserve the high-frequency components of the image. DWT separates the image into different sub band images, namely, LL, LH, HL, and HH. A high-frequency sub band contains the high frequency component of the image. The interpolation can be applied to these four sub band images. In the wavelet domain, the low-resolution image is obtained by low-pass filtering of the high-resolution image. The low resolution image (LL sub band), without quantization (i.e., with double-precision pixel values) is used as the input for the proposed resolution enhancement process. In other words, low frequency sub band images are the low resolution of the original image. Therefore, instead of using low-frequency sub band images, which contains less information than the original input image, we are using this input image through the interpolation process. Hence, the input low-resolution image is interpolated with the half of the interpolation factor, $\alpha/2$, used to interpolate the high-frequency sub-bands, as shown inFig.2 In order to preserve more edge information, i.e., obtaining a sharper enhanced image, we have proposed an intermediate stage in high frequency sub band interpolation process.

As shown in Fig.1, the low-resolution input satellite image and the interpolated LL image with factor 2 are highly correlated. The difference between the LL sub band image and the low-resolution input image are in their high-frequency components. Hence, this difference image can be use in the intermediate process to correct the estimated high-frequency components. This estimation is performed by interpolating the high-frequency sub bands by factor 2 and then including the difference image (which is high-frequency components on low-resolution input image) into the estimated high-frequency images, followed by another interpolation with factor $\alpha/2$ in order to reach the required size for IDWT process.

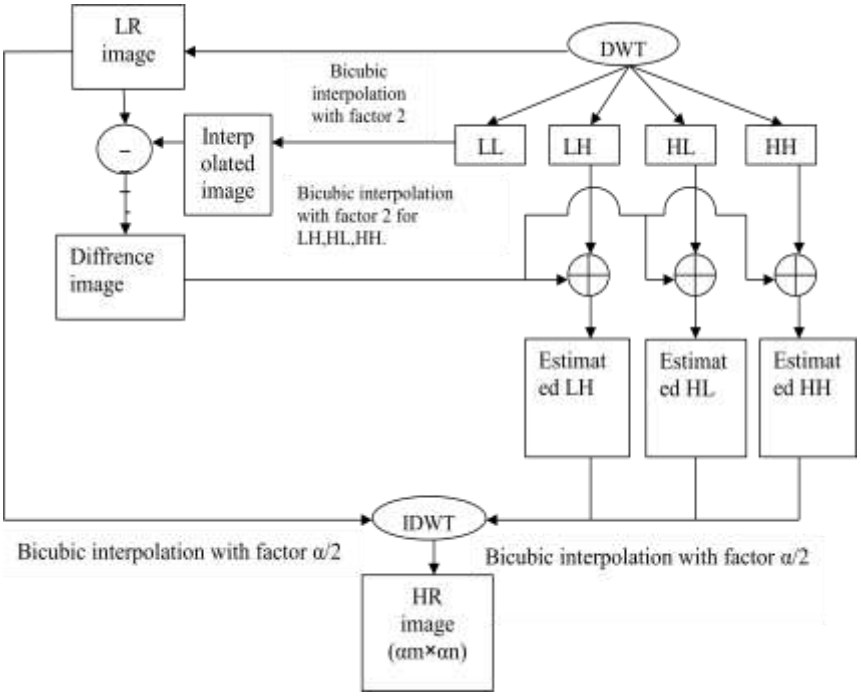

Fig.1 Conventional enhancement block diagram

IV.PROPOSED METHOD

A high-frequency sub band contains the high frequency component of the image. In other words, low frequency sub band images are the low resolution of the original image.Therefore, instead of using low-frequency sub band images, which contains less information than the original input image. we are using this input image through the interpolation process. IWT also separates the image into different sub band images, namely, LL, LH, HL, and HH. A high-frequency sub band contains the high frequency component of the image. Hence, the input low-resolution image is interpolated with the half of the interpolation factor α/2, used to interpolate the high frequency sub bands, as shown in Fig.2.In order to preserve more edge information, i.e., obtaining a sharper enhanced image we have proposed an intermediate stage in high frequency sub band interpolation process. The DWT's LH sub band and IWT's LH sub band are added, and then interpolated by factor 2 to form estimated LH. The same procedure is repeated for HL and HH to form estimated HL and estimated HH respectively. In this process we are not considering LL sub bands of both DWT and IWT. In the place of LL we are taking input low resolution image. After estimation input low resolution image and high-frequency sub bands are interpolated by factor $\alpha/2$ and to reach the required size for IDWT process.

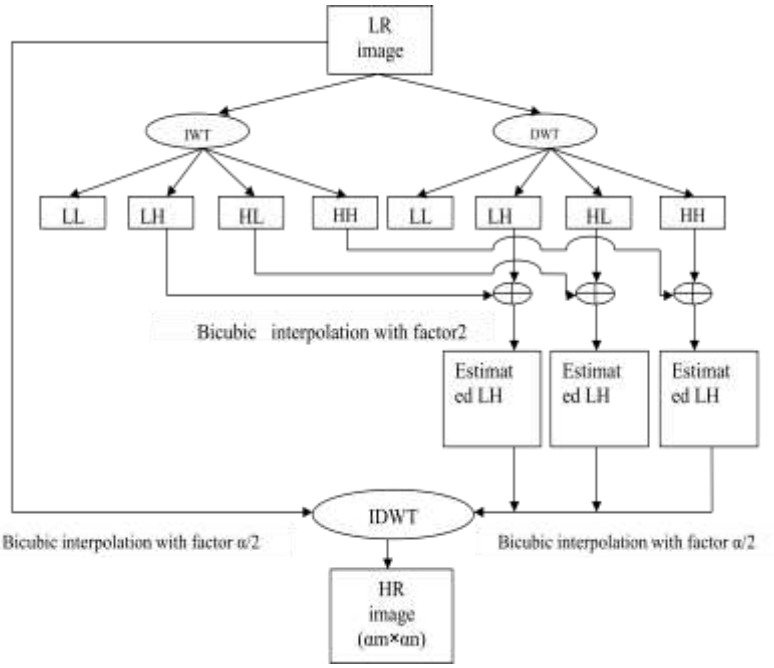

Fig.2 Proposed method Block Diagram

V.CONCLUSION

Satellite image enhancement is area of concern in the field of digital image processing. In literature many conventional works has been reported but fails to achieve desired enhancement results based on PSNR and MSE parameters based on superiority of the output images. In this paper a new enhancement technique has been reported based on the interpolation of the high frequency sub band images by DWT, IWT and the input low resolution image. The PSNR improvement of the proposed technique is upto 2.19 dB compared with conventional method.

VI.ACKNOWLEDGMENT

The authors would like to thank Dr. A.Temizel from Bilkent University for providing the output of the WZP resolution enhancement technique [14]. Moreover, the authors would like to acknowledge Prof. Dr. I. Selesnick from Polytechnic University for providing the DWT codes in MATLAB. Furthe rmore, the authors would like to thank Google Earth and Satellite Imaging Corporation for providing satellite images for research purposes.

VII.SIMULATION RESULTS

(a)

(d)

(b)

(e)

(c)

(f)

(a) low resolution image, (b) Bilinear image (c) Bicubic image (d) Super resolution image using WZP(e) DWT based image resolution enhancement output, (f) proposed method output image.

In the above images the proposed method output image superiority is better than the input image and all conventional methods output images. Hence it is called best enhancement method for low resolution satellite images.

Enhancement Method	PSNR	MSE
Bi-Linear	25.7841	0.4247
Bi-cubic	29.7570	0.2697
WZP	35.6436	0.0695
DWT	35.6700	0.0690
Proposed	37.6488	0.0438

TABULAR COLUMN: PSNR and MSE Results for resolution enhancement from 128×128 to 512×512 (α=4) for the proposed technique compared with conventional resolution enhancement techniques.

VIII.REFERENCES

[1]Hasan Demirel and Gholamreza Anbarjafari, "Discrete Wavelet Transform-Based Satellite Image Resolution Enhancement", IEEE transactions on geosciences and remote sensing, June 2013.

[2]Integer Wavelet transform using the lifting scheme.Geert uyterhoeven,Dirk Roose, adhemar, Bultheel,Department of computer science.

[3] H. Demirel, G. Anbarjafari, and S. Izadpanahi, "Improved motion-based localized super resolution technique using discrete wavelet transform for low resolution video enhancement," in *Proc. 17th EUSIPCO*, Edinburgh, U.K., Aug. 2009, pp. 1097–1101.

[4] L. Yi-bo, X. Hong, and Z. Sen-yue, "The wrinkle generation method for facial reconstruction based on extraction of partition wrinkle line features and fractal interpolation," in *Proc. 4th ICIG*, Aug. 22–24, 2007, pp. 933–937.

[5] Y. Rener, J. Wei, and C. Ken, "Downsample-based multiple description coding and post-processing of decoding," in *Proc. 27th CCC*, Jul. 16–18, 2008, pp. 253–256.

[6] C. B. Atkins, C. A. Bouman, and J. P. Allebach, "Optimal image scaling using pixel classification," in *Proc. ICIP*, Oct. 7–10, 2001, vol. 3, pp. 864–867.

[7] Y. Piao, L. Shin, and H. W. Park, "Image resolution enhancement using inter-subband correlation in wavelet domain," in *Proc. IEEE ICIP*, 2007,

[8] G. Anbarjafari and H. Demirel, "Image super resolution based on interpolation of wavelet domain high frequency subbands and the spatial domain input image," *ETRI J.*, vol. 32, no. 3, pp. 390–394, Jun. 2010.

[9] W. K. Carey, D. B. Chuang, and S. S. Hemami, "Regularity-preserving image interpolation," *IEEE Trans. Image Process.*, vol. 8, no. 9, pp. 1295–1297, Sep. 1999.

[10] X. Li and M. T. Orchard, "New edge-directed interpolation," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1521–1527, Oct. 2001.

[11] K. Kinebuchi, D. D. Muresan, and T.W. Parks, "Image interpolation using wavelet based hidden Markov trees," in *Proc. IEEE ICASSP*, 2001, vol. 3, pp. 7–11.

[12] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical

signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.
[13] S. Zhao, H. Han, and S. Peng, "Wavelet domain HMT-based image super resolution," in *Proc. IEEE ICIP*, Sep. 2003, vol. 2, pp. 933–936.
[14] A. Temizel and T. Vlachos, "Image resolution upscaling in the wavelet domain using directional cycle spinning," *J. Electron. Imaging*, vol. 14, no. 4, p. 040501, 2005.
[15] A. Gambardella andM.Migliaccio, "On the superresolution of microwave scanning radiometer measurements," *IEEE Geosci. Remote Sens. Lett.* vol. 5, no. 4, pp. 796–800, Oct. 2008.
[16] V. A. Tolpekin and A. Stein, "Quantification of the effects of land-coverclass spectral separability on the accuracy of Markov-random-field-based superresolution mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 9, pp. 3283–3297, Sep. 2009.
[17] A. Temizel and T. Vlachos, "Wavelet domain image resolution enhancement using cycle-spinning," *Electron. Lett.*, vol. 41, no. 3, pp. 119–121, Feb. 3, 2005.
[18] L. A. Ray and R. R. Adhami, "Dual tree discrete wavelet transform with application to image fusion," in *Proc. 38th Southeastern Symp. Syst. Theory*, Mar. 5–7, 2006, pp. 430–433.
[19] A. Temizel, "Image resolution enhancement using wavelet domain hidden Markov tree and coefficient sign estimation," in *Proc. ICIP*, 2007, vol. 5, pp. V-381–V-384.
[20] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 2007.

International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 77 - 81, 2015, ISSN:- 2319–8656

Improved Firefly Algorithm for Unconstrained Optimization Problems

Khalil AL-Wagih*
Faculty of Computer science & Information System,
Thamar University, Thamar, Republic of Yemen
E-mail: khalilwagih@gmail.com

Abstract: in this paper, an improved firefly algorithm with chaos (IFCH) is presented for solving unconstrained optimization problems. Several numerical simulation results show that the algorithm offers an efficient way to solve unconstrained optimization problems, and has a high convergence rate, high accuracy and robustness.

Keywords: Firefly algorithm; metaheuristic; optimization; chaos; unconstrained optimization

.

1. INTRODUCTION

The problem of finding the global optimum of a function with large numbers of local minima arises in many scientific applications. In typical applications, the search space is large and multi-dimensional. Many of these problems cannot be solved analytically, and consequently, they have to be addressed by numerical algorithms. Moreover, in many cases, global optimization problems are non-differentiable. Hence, the gradient-based methods cannot be used for finding the global optimum of such problems. To overcome these problems, several modern heuristic algorithms have been developed for searching near-optimum solutions to the problems. These algorithms can be classified into different groups, depending on the criteria being considered, such as population-based, iterative based, stochastic, deterministic, etc. Depending on the nature of the phenomenon simulated by the algorithms, the population-based heuristic algorithms have two important groups: Evolutionary Algorithms (EA) and swarm intelligence based algorithms.

Some of the recognized evolutionary algorithms are: Genetic Algorithms (GA) [1], Differential Evolution (DE) [2] and [3], Evolution Strategy (ES) [4] and Artificial Immune Algorithm (AIA) [5] etc. Some of the well known swarm intelligence based algorithms are: Particle Swarm Optimization (PSO) , Ant Colony Optimization (ACO) , Shuffled Frog Leaping (SFL) , and Artificial Bee Colony (ABC) algorithms , etc. Besides the evolutionary and swarm intelligence based algorithms, there are some other algorithms which work on the principles of different natural phenomena. Some of them are: the Harmony Search (HS) algorithm, the Gravitational Search Algorithm (GSA) , Biogeography-Based Optimization (BBO) , the Grenade Explosion Method (GEM) , the league championship algorithm and the charged system search .

This paper is organized as follows: after introduction, the original firefly algorithm is briefly introduced in section 2. In section 3, the proposed algorithm is described, while the results are discussed in section 4. Finally, conclusions are presented in section 5.

2. FIREFLY ALGORITHM

The Firefly Algorithm [FA] is one of many new optimization techniques that have been proposed over the past years. It was proposed by Yang in 2009 [6] and it has since then been applied in several applications bbecause of its few parameters to adjust, easy to understand, realize, and compute, it was applied to various fields, such as codebook of vector quantization [7], in-line spring-mass systems [8]; mixed variable structural optimization [9]; nonlinear grayscale image enhancement [10], travelling salesman problems [11], continuously cast steel slabs [12], promoting products online [13], nonconvex economic dispatch problems [14], chiller loading for energy conservation [15], stock market price forecasting [16], and multiple objectives optimization [17]. Although the algorithm has many similarities with other swarm based algorithms such as Particle Swarm Optimization [18], Artificial Bee Colony Optimization [19]and Ant Colony Optimization [6], the FA has proved to be much simpler both in concept and implementation and has better performance compared to the other techniques.

2.1 Flashing behaviour of Fireflies

The FA was based on the flashing patterns and behaviour patters of the fireflies. The fireflies use the flashing patterns to communicate with each other. Yang did not mimic their behaviour in full detail, but created a simplified algorithm based on the following three rules:

i. All fireflies are unisexual, so that one firefly will be attracted to other fireflies regardless of their sex;
ii. Attractiveness is proportional to the firefly's brightness; for any couple of flashing fireflies, the less bright one will move towards the brighter one; attractiveness is proportional to the brightness which decreases with increasing distance between fireflies; if there are no

brighter fireflies than a particular firefly, this individual will move randomly in the space;

iii. The brightness of a firefly is somehow related to the analytical form of a cost function; for a maximization problem, brightness can be proportional to the value of the cost function; other forms of brightness can be defined in a similar matter to the fitness function in genetic algorithms.

2.2 Attractiveness and Light Intensity

In the algorithm, two important factors are involved: the variation of light intensity and the formulation of the attractiveness. For example, suppose that the attractiveness of a firefly is determined by its brightness, which in turn is associated with the encoded objective function, then the higher of the brightness and, the better the location and the more fireflies will be attracted to the direction. However, if the brightness is equal, the fireflies will move randomly. As light intensity and thus attractiveness decreases as the distance from the source increases, the variations of light intensity and attractiveness should be monotonically decreasing functions.

In order to implement FA, there are some definitions:

Definition 1: the variation of light intensity;

We know, the light intensity varies according to the inverse square law

$$I(r) = {}^{I_s}/_{r^2} \quad (1)$$

Where I(r) is the light intensity at a distance r and I_s is the intensity at the source.

When the medium is given, the light intensity can be determined as follows:

$$I(r) = I_0 e^{-\gamma r} \quad (2)$$

To avoid the singularity at r=0 in (1), the equations can be approximated in the following Gaussian form:

$$I(r) = I_0 e^{-\gamma r^2} \quad (3)$$

Where γ is light absorption coefficient.

Definition 2: *formulation of the attractiveness*

As firefly attractiveness is proportional to the light intensity seen by adjacent fireflies, we can now define the attractiveness β of a firefly by

$$\beta = \beta_0 e^{-\gamma r^2} \quad (4)$$

Where β_0 is the attractiveness at r=0.

Definition 3: *formulation of location moving*

$$x_i(t+1) = x_i(t) + \beta\left(x_j(t) - x_i(t)\right) + \alpha\varepsilon_i \quad (5)$$

Where $x_i(t+1)$is the position of x_i after *t+1* times movements; α is the step parameter which varies between [0,1] ; $\varepsilon_i$is a random factor conforming Gaussian distribution between[0,1].

The basic steps of the FA are summarized as the pseudo code shown in Fig. 1 which consists of the three rules discussed above.

firefly algorithm
Begin
Objective function $f(x)$, $x = (x_1, \ldots, x_d)^T$
Generate initial population of $n$fireflies x_i, $i = 1, 2, \ldots, n$
Formulate light intensity Iso that it is associated with $f(x)$
While *(t<MaxGeneration)*
Define absorption coefficient γ
for *i = 1 : n($n$fireflies)*
for *j = 1 : n($n$fireflies)*
if *(I_j>I_i),*
move firefly $i$towards j
end if
Vary attractiveness with distance r via $e^{-\gamma r^2}$
Evaluate new solutions and update light intensity
end for *j*
end for *i*
Rank the fireflies and find the current best
end while
Post-processing the results and visualization
End

Fig. 1 Pseudo code of the firefly algorithm

3. THE PROPOSED ALGORITHM (IFCH) FOR UNCONSTRAINED OPTIMIZATION PROBLEMS

Generating random sequences with a long period, and a good consistency is very important for easily simulating complex phenomena, sampling, numerical analysis, decision making and especially in heuristic optimization [20]. Its quality determines the reduction of storage and computation time to achieve the desired accuracy [21]. Chaos is a deterministic, random-like process found in nonlinear, dynamical system, which is non-period, non-converging and bounded. Moreover, it depends on its initial condition and parameters [22-24]. Applications of chaos in several disciplines including operations research, physics, engineering, economics, biology, philosophy and computer science[25-27].

Recently chaos is extended to various optimization areas because it can more easily escape from local minima and improve global convergence in comparison with other stochastic optimization algorithms [28-34]. Using chaotic sequences in FireflyAlgorithm can be helpfully improve the reliability of the global optimality, and also enhance the quality of the results.

In the proposed chaotic Firefly Algorithm, we used chaotic maps to tune the Firefly Algorithm parameters and improve the performance [20]. The steps of the proposed chaotic firefly algorithm for solving definite integral are as follows:

Step 1 Generate the initial population of fireflies,

$\{x_1, x_2, x_3, \ldots, x_n\}$

Step 2 Compute intensity for each firefly member,$\{I_1, I_2, I_3, \ldots, I_n\}$

Step 3 Calculate the parameters (β , γ) using the following Sinusoidal map[35]:

$$Y_{n+1} = \cos(k\cos^{-1}(Y_n)) \; Y \in (-1,1)$$

where n is the iteration number.

Step 4 Move each firefly x_i towards other brighter fireflies. The position of each firefly is updated by

$$x_i(t+1) = x_i(t) + \beta_0 e^{-\gamma r^2}\left(x_j(t) - x_i(t)\right) + \alpha\varepsilon_i$$

Where α computed by the following randomness equation as shown below:

$$\alpha^i = \alpha_{max} - (\alpha_{max} - \alpha_{min})\left(\frac{I^i_{max} - I^i_{mean}}{I^i_{max} - I^i_{min}}\right) \qquad (6)$$

In this equation α^i represents randomness parameters at cycle i. α_{max} and α_{min} represent maximum and minimum randomness parameters defined in the algorithm respectively. I^i_{max} and I^i_{min} represent maximum light intensity, minimum light intensity and mean value of light intensity of all fireflies at cycle i respectively.

Step 5 Update the solution set.

Step 6 Terminate if a termination criterion is fulfilled; otherwise go to Step 2.

4. EXPERIMENTAL RESULTS

Six well known test functions have been given to verify the weight of the proposed algorithm. The initial parameters are set at *n*= 40; maximum iteration number = 100;$\alpha_{max} = 0.8$; $\alpha_{min} = 0.1$. The results of IFCH algorithm are conducted from 50 independent runs for each problem. The selected chaotic map for all problems is the Sinusoidal map for β ,γ values, and randomized for α values, whose equations is shown above.

All the experiments were performed on a Windows 7 Ultimate 64-bit operating system; processor Intel Core i7 760 running at 2.81 GHz; 8GB of RAM and code was implemented in C#.

Test problems are considered to extensively investigate the performance of the IFCH algorithm, and they are presented as follows:

The first is Sphere function, defined as

$$\text{Min } f_1 = \sum_{i=1}^{N} x_i^2$$

Where global optimum x*= (0,0,…,0) and f(x*) = 0 for -100 $\le x_i \le$100.

The second is Rosenbrock function, defined as

$$\text{Min } f_2 = \sum_{i=1}^{N-1} (100\ (x_{i+1} - x_i^i)^2 + (x_i - 1)^2)$$

Where global optimum x*= (1,1,…,1) and f(x*) = 0 for -100 $\le x_i \le$100.

The third is generalized Rastrigrin function, defined as

$$\text{Min } f_3 = \sum_{i=1}^{N} (x_i^2 - 10\ \cos(2\pi x_i) + 10)$$

Where global optimum x*= (0,0,…,0) and f(x*) = 0 for -10 $\le x_i \le$10.

The fourth function is as follows:

$$\text{Min } f_4 = \sum_{i=1}^{N} z_i^2 - 450$$

The fifth is generalized Griewank function, defined as

$$\text{Min } f_5 = \frac{1}{4000}\sum_{i=1}^{N} x_i^2 - \prod_{i=1}^{N} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$$

Where global optimum x*= (0,0,…,0) and f(x*) = 0 for -600 $\le x_i \le$600.

The sixth is Schwefel's function, defined as

$$\text{Min } f_6 = \sum_{i=1}^{N} |x_i| + \prod_{i=1}^{N} |x_i|$$

Where global optimum x*= (0,0,…,0) and f(x*) = 0 for -100 $\le x_i \le$100.

Table 1 the solution of proposed algorithm and firefly algorithm

Test Problem	Algorithm	Best	Worst	Mean	Standard Deviation
f1	FA	6.0076e+001	2.5205e+002	1.7603e+002	4.4359e+001
	IFCH	2.2757e-010	7.0950e-009	1.8091e-009	1.9488e-009
f2	FA	9.6629e+004	8.9231e+005	3.3964e+005	1.9909e+005
	IFCH	2.4458e-001	8.1098e+003	5.3162e+002	1.7208e+003
f3	FA	2.9511e+001	5.2810e+001	4.2476e+001	5.7278e+000
	IFCH	3.7993e-009	9.9496e-001	3.3166e-002	1.8165e-001
f4	FA	-3.5043e+002	-1.6889e+002	-2.5642e+002	4.4840e+001
	IFCH	-4.5000e+002	-4.5000e+002	-4.5000e+002	1.2822e-009
f5	FA	2.0566e+000	3.2913e+000	2.5946e+000	3.2134e-001
	IFCH	4.4744e-010	1.3045e-001	3.7959e-002	3.8783e-002
f6	FA	3.1353e+001	5.0629e+001	4.2350e+001	4.9070e+000
	IFCH	3.6976e-005	2.2687e-004	8.1858e-005	3.7660e-005

5. CONCLUSIONS

This paper introduced an improved Firefly Algorithm by blending with chaos for unconstrained optimization problems. The proposed algorithm employs a novel method for generating new solutions that enhances accuracy and convergence rate of FA. The proposed algorithm has been successfully applied to various benchmarking of unconstrained optimization problems. Case study results reveal that the proposed algorithm can find the global optimal solutions and is a powerful search algorithm for various unconstrained optimization problems.

.

REFERENCES

[1] H., Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, USA (1975).

[2] R.Storn, and K.Price, ‘‘Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces’’, J. Global Optim., 11, pp. 341–359 (1997).

[3] K.Price, , R. Storn, and A.Lampinen, , Differential Evolution—A Practical Approach to Global Optimization, Springer-Verlag, Berlin, Germany (2005).

[4] T.P.Runarsson, and X. Yao, ‘‘Stochastic ranking for constrained evolutionary optimization’’, IEEE Trans. Evol. Comput., 4(3), pp. 284 294(2000).

[5] L.J. Fogel, , A.J. Owens, and M.J.Walsh, , Artificial Intelligence Through Simulated Evolution, John Wiley, New York, USA (1966).

[6] X.-S. Yang, *Nature-inspired metaheuristic algorithms*: Luniver Press, 2010.

[7] M.-H. Horng and T.-W. Jiang, "The codebook design of image vector quantization based on the firefly algorithm," in *Computational Collective Intelligence. Technologies and Applications*, ed: Springer, 2010, pp. 438-447.

[8] R. Dutta, R. Ganguli, and V. Mani, "Exploring isospectral spring–mass systems with firefly algorithm," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, vol. 467, pp. 3222-3240, 2011.

[9] A. H. Gandomi, X.-S. Yang, and A. H. Alavi, "Mixed variable structural optimization using firefly algorithm," *Computers & Structures*, vol. 89, pp. 2325-2336, 2011.

[10] T. Hassanzadeh, H. Vojodi, and F. Mahmoudi, "Non-linear grayscale image enhancement based on firefly algorithm," in *Swarm, Evolutionary, and Memetic Computing*, ed: Springer, 2011, pp. 174-181.

[11] G. K. Jati, "Evolutionary discrete firefly algorithm for travelling salesman problem," in *Adaptive and Intelligent Systems*, ed: Springer, 2011, pp. 393-403.

[12] O. K. K. L. JEKLENE, "Optimization of the Quality of Continuously Cast Steel Slabs Using the Firefly Algorithm," *Materiali in tehnologije*, vol. 45, pp. 347-350, 2011.

[13] H. Banati and M. Bajaj, "Promoting products online using firefly algorithm," in *Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on*, 2012, pp. 580-585.

[14] X.-S. Yang, S. S. Sadat Hosseini, and A. H. Gandomi, "Firefly algorithm for solving non-convex economic dispatch problems with valve loading effect," *Applied Soft Computing*, vol. 12, pp. 1180-1186, 2012.

[15] L. d. S. Coelho and V. C. Mariani, "Improved firefly algorithm approach for optimal chiller loading for energy conservation," *Energy and Buildings*, 2012.

[16] A. Kazem, E. Sharifi, F. K. Hussain, M. Saberi, and O. K. Hussain, "Support vector regression with chaos-based firefly algorithm for stock market price forecasting," *Applied Soft Computing*, 2012.

[17] X.-S. Yang, "Multiobjective firefly algorithm for continuous optimization," *Engineering with Computers*, pp. 1-10, 2013.

[18] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," *Swarm intelligence*, vol. 1, pp. 33-57, 2007.

[19] P.-W. TSai, J.-S. Pan, B.-Y. Liao, and S.-C. Chu, "Enhanced artificial bee colony optimization," *International Journal of Innovative Computing, Information and Control*, vol. 5, pp. 5081-5092, 2009.

[20] B. Alatas, "Chaotic harmony search algorithms," *Applied Mathematics and Computation*, vol. 216, pp. 2687-2699, 2010.

[21] W. Gong and S. Wang, "Chaos Ant Colony Optimization and Application," in *Internet Computing for Science and Engineering (ICICSE), 2009 Fourth International Conference on*, 2009, pp. 301-303.

[22] B. Alatas, "Chaotic bee colony algorithms for global numerical optimization," *Expert Systems with Applications*, vol. 37, pp. 5682-5687, 2010.

[23] A. Gandomi, X.-S. Yang, S. Talatahari, and A. Alavi, "Firefly algorithm with chaos," *Communications in Nonlinear Science and Numerical Simulation*, vol. 18, pp. 89-98, 2013.

[24] J. Mingjun and T. Huanwen, "Application of chaos in simulated annealing," *Chaos, Solitons & Fractals*, vol. 21, pp. 933-941, 2004.

[25] L. d. S. Coelho and V. C. Mariani, "Use of chaotic sequences in a biologically inspired algorithm for engineering design optimization," *Expert Systems with Applications*, vol. 34, pp. 1905-1913, 2008.

[26] M. S. Tavazoei and M. Haeri, "Comparison of different one-dimensional maps as chaotic search pattern in chaos optimization algorithms," *Applied Mathematics and Computation*, vol. 187, pp. 1076-1085, 2007.

[27] R. Hilborn, *Chaos and nonlinear dynamics: an introduction for scientists and engineers*: oxford university press, 2000.

[28] D. He, C. He, L.-G. Jiang, H.-W. Zhu, and G.-R. Hu, "Chaotic characteristics of a one-dimensional iterative map with infinite collapses," *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, vol. 48, pp. 900-906, 2001.

[29] A. Erramilli, R. Singh, and P. Pruthi, *Modeling packet traffic with chaotic maps*: Citeseer, 1994.

[30] R. M. May, "Simple mathematical models with very complicated dynamics," in *The Theory of Chaotic Attractors*, ed: Springer, 2004, pp. 85-93.

[31] A. Wolf, "Quantifying chaos with Lyapunov exponents," *Chaos*, pp. 273-290, 1986.

[32] R. L. Devaney, "An introduction to chaotic dynamical systems," 2003.
[33] C. Letellier, *Chaos in nature* vol. 81: World Scientific Publishing Company, 2013.
[34] R. Barton, "Chaos and fractals," *The Mathematics Teacher*, vol. 83, pp. 524-529, 1990.
[35] I. Fister, I. JrFister, X.-S. Yang, and J. Brest, "A comprehensive review of firefly algorithms," *Swarm and Evolutionary Computation*, 2013.

International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 82 - 85, 2015, ISSN:- 2319–8656

Human Perception and Recognition of Musical Instruments: A Review

Satish Ramling Sankaye
MGM Dr. G.Y. Pathrikar College of CS and IT,
Aurangabad, India

U. S. Tandon
Department of Physics,
College of Natural and Computational Sciences,
Haramaya University,
Dire Dawa, Ethiopia

Abstract: Musical Instrument is the soul of music. Musical Instrument and Player are the two fundamental component of Music. In the past decade the growth of a new research field targeting the Musical Instrument Identification, Retrieval, Classification, Recognition and management of large sets of music is known as Music Information Retrieval. An attempt to review the methods, features and database is done.

Keywords: Musical Instrument; Monophonic; Polyphonic; Classification Data Model;

1. INTRODUCTION

Human perception in musical applications is especially important, since musical sounds are designed merely for human audition. The study of music signal is useful in teaching and evaluation of music. The human vocal apparatus which generates speech also generates music. Therefore, the studies reveal similarities in the spectral and temporal properties of music and speech signals. Hence, many techniques developed to study speech signal are employed to study music signals as well.

To make music, two essential components are needed: the player and the instrument. Hence one of the key aspects in the research of Music has focused on the internal contents of music, the Musical Instrument. In the past decade the growth of a new research field targeting the Musical Instrument Identification, Retrieval, Classification, Recognition and management of large sets of music is known as Music Information Retrieval. Musical instrument Identification is edged on classification of single note (Monophonic), more than one instrument notes at a time (Polyphonic), distinction of instruments in continuous recording or Classification of family/genre. Musical instruments are classified into five families depending on the sound produced as percussion, brass, string, woodwind and keyboard [4], [7].

Table 1: The musical instrument collection

Family	Instruments
Brass	French horn, Trombone, Trumpet, Tuba
Keyboard	Piano, Harmonium
Percussion	Bell, Bongo, Chime, Conga, Cymbal, Dholki, Drum, Gong, Tambourine, Triangle, Timbales, Tympani, Tabla,
String	Guitar, Violin, Sitar, Vichitraveena, Saraswativeena, Rudraveena
Woodwind	Shehnai, Oboe, Saxophone, Flute

The paper is organized as follows: Section 2 describes the different databases studied. The different classification and pattern recognition techniques are discussed in section 3. Finally section 4 furnishes the conclusion.

2. DATABASE

Musical Instrument Identification leads to the aspect of initially recording the sound sample from different sources. It can be recorded directly while playing the instrument by using tape recorder, mobile or any other electronic gadget meant for sound recording in natural environment. Also for the study purpose, the musical instruments are played in an anechoic room at a professional studio. Some of the commonly used databases studied by most of the researchers include:

2.1 Musical Audio Signal Separation (MASS) Dataset

This database was created to help to evaluate of Musical Audio Signal Separation algorithms and statements on a representative set of professionally produced music (i.e. real music recordings). It included several song snips of a few seconds (10s-40s) with the following contents:

- *Tracks (with/without effects):* Stereo Microsoft PCM WAV files (44.1Khz, 24 bits) of every instrumental track including and/or without including effects (plugins enabled or disabled in the project file used for production)
- *Description of the effects*: When available, included a description of the plugins used to modify the tracks without effects.
- *Lyrics:* When available, lyrics are included.

The dataset was compiled by M. Vinyes (MTG former member). Bearlin and Sargon have released the tracks of their songs and Sergi Vila at Garatge Productions and Juan Pedro Barroso at Kcleta Studios [14]. It is available online *www.sargonmetal.com.*

2.2 University of Iowa musical instrument samples

The Musical Instrument Samples Database has been divided into two categories: pre-2012 and post-2012 files. The pre-2012 files are the original sound files that are present on website http://theremin.music.uiowa.edu/MIS.html [17] since the end of May 2014. These sound files were recorded in the Anechoic Chamber at the Wendell Johnson Speech and Hearing Center as early as 1997. This category consists of

mono files for woodwinds, brass, and string instruments at a 16-bit, 44.1 kHz format with Neumann KM 84 Microphones. It also contains stereo files for the most recent recordings of string instruments done between December 2011 and May 2012, and percussion instruments done between March and June 2013 at a 24-bit, 44.1 kHz format with 3 Earthworks QTC40 microphones in a Decca Tree configuration.

The post-2012 files are experimental sound files. They are edited sound files extracted from the University of Iowa Electronic Music Studios website from the pre-2012 category. Each instrument from the string, woodwind, brass, and percussion families, excluding the guitar, piano, and Soundmines folder, has been edited as of July 24, 2014 for public and research use. All files from the string, woodwind, brass, and percussion families have been converted to a 24-bit, 44.1 kHz stereo format. Whenever possible, mid-side processing was applied to these files to widen the stereo field. These files were created in Studio 1 of the University Electronic Music Studios in the Becker Communication Studies

2.3 McGill university master samples

The first release of McGill University Master Samples [MUMS] (Opolko & Wapnick, 1987) featured 3 CDs of recorded, high quality instrument samples. Recently, the library has been expanded to 3 DVDs (Opolko & Wapnick, 2006) and contains samples of most standard classical, some non-standard classical, and many popular music instruments. There are 6546 sound samples in the library, divided between string (2204), keyboard (1595), woodwind (1197), percussion (1087, out of which 743 are non-pitched), and brass (463) families. In principle, each note of each instrument has been recorded separately (44.1 kHz, 24-bit), and most instruments feature several articulation styles. Typically there are 29 samples per instrument, which means that the whole pitch range of the available instruments is not consistently covered. The coverage is nevertheless impressive.

This library is one of the most often used sources of instrument samples within instrument recognition and classification research, sound synthesis and manipulation studies. The library has also been the source for an edited database (SHARC) of steady-state instrument spectra.

2.4 Real World Computing Music Database:

RWC [13] Music Database comprises of four original component Databases. Popular Music Database (100 pieces), Royalty-Free Music Database (15 pieces), Classical Music Database (50 pieces), and Jazz Music Database (50 pieces). Recently two more Database component viz Music Genre Database (100 pieces) and Musical Instrument Sound Database (50 instruments) are added.

The Database of Musical Instrument Sound covers 50 musical instruments and provides, in principle, three variations for each instrument. In all about 150 performances of different instruments are present. To provide a wide variety of sounds, following approach has been taken.

- *Variations (3 instrument manufacturers, 3 musicians):* Each variation featured, in principle, an instrument from a different manufacturer played by a different musician.
- *Playing style (instrument dependent):* Within the range possible for each instrument, many playing styles have been recorded.
- *Pitch (total range):* For each playing style, the musician played individual sounds at half-tone intervals over the entire range of tones that could be produced by that instrument.
- *Dynamics (3 dynamic levels):* Recording was also done for each playing style at three levels of dynamics (forte, mezzo, piano) spanning the total range of the instrument.

The sounds of these 50 instruments were recorded at 16 bit / 44.1 kHz and stored in 3544 monaural sound files having a total size of about 29.1 GBytes and a total playback time (including mute intervals) of about 91.6 hours [13].

3. CLASSIFCATION DATA MODEL

Various Features of Musical Signal have been studied, which are classified as Temporal, Spectral, Time-Domain, and Frequency Domain. Note onset detection and localization is also useful for a number of analysis and indexing techniques for musical signals [2]. Attack, Decay, Sustain and Release [2], [8] are other important features of sound waveform's energy distribution. After studying these feature set the different model are being implemented on the feature set for the Identification of the musical instrument or classifying the excerpt as a member of particular family. The various commonly studied models are discussed below:

3.1 Support Vector Machines

Support Vector Machine (SVM) [15] is a supervised learning method that belongs to a family of linear classifiers used for classification and regression. However, SVM is closely related to neural networks. It is based on some relatively simple ideas but constructs models that are complex enough and it can lead to high performances in real world applications.

The basic idea behind Support Vector Machines is that it can be thought of as a linear method in a high-dimensional feature space nonlinearly related to input space. Therefore in practice it does not involve any computation in the high-dimensional space. All necessary computations are performed directly in input space by the use of kernels. Therefore the complex algorithms for nonlinear pattern recognition, regression, or feature extraction can be used pretending that the simple linear algorithms are used.

The key to the success of SVM is the kernel function which maps the data from the original space into a high dimensional (possibly infinite dimensional) feature space. By constructing a linear boundary in the feature space, the SVM produces non-linear boundaries in the original space. When the kernel function is linear, the resulting SVM is a maximum-margin hyperplane. Given a training sample, a maximum-margin hyperplane splits a given training sample in such a way that the distance from the closest cases (support vectors) to the hyperplane is maximized. Typically, the number of support vectors is much less than the number of the training sample. Nonlinear kernel functions such as the polynomial kernel and the Gaussian (radial basis function) kernel are also commonly used in SVM. One of the most important advantages for the SVM is that it guarantees generalization to some extent. The decision rules reflect the regularities of the training data rather than the incapabilities of the learning machine. Because of the many nice properties of SVM, it has been widely applied to virtually every research field.

3.2 Hidden Markov Model

A hidden Markov model (HMM) [15] is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved state. An HMM can be considered as the simplest dynamic Bayesian network. Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges and bioinformatics.

The main characteristic of Hidden Markov Model is that it utilizes the stochastic information from the musical frame to recognize the pattern. In a hidden Markov model, the state is not directly visible, but the dependence of output on the state is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states.

Hidden Markov Models are widely used as general-purpose speech recognition and musical instrument as well as music identification systems. The basic reason why HMMs are used in music/speech recognition is that a music/speech signal could be viewed as a piecewise stationary signal or a short-time stationary signal.

Another reason why HMMs are popular is that they can be trained automatically and they are simple and computationally feasible to use.

3.3 Gaussian Mixture Model

A Gaussian Mixture Model (GMM) was used as classification tool [10], [15]. GMMs belong to the class of pattern recognition systems. They model the probability density function of observed variables using a multivariate Gaussian mixture density. Given a series of inputs, it refines the weights of each distribution through expectation-maximization algorithms.

In order to construct the models for the music recognition system, they calculated the features for all samples of the database and store the features for each class separately. Then, a Gaussian Mixture Model (GMM), θ_i, for each class is built (i.e., with i = 1...C, where C denotes the number of different classes), using a standard Expectation Maximization (EM) algorithm. EM algorithm is initialized by a deterministic procedure based on the Gaussian means algorithm. A new song is classified into a new category by computing the likelihood of its features given in the classification models, θ_i, with i = 1...C. Summing up these likelihood values, the song is assigned to the class that has the maximum summation value [11].

3.4 Probabilistic latent component analysis (PLCA)

Probabilistic Latent Component Analysis (PLCA) or Non-negative Matrix Factorization (NMF) is efficient frameworks for decomposing the mixed signal into individual contributing components [1]. In NMF approach, the features representing each instrument are the spectral dictionaries which are used to decompose the polyphonic spectra into the source instruments. PLCA interprets this task probabilistically by assuming the spectrum to be generated from an underlying probability density function (pdf), and estimates the joint distribution of observed spectra and a set of underlying latent variables.

Probabilistic Latent Component Analysis [1] is based on modelling the normalized magnitude of the observed spectrum V (f, t) as the probability distribution Pt(f) at time frame index t and frequency bin index f. Pt(f) is factorized into many latent components as

$$P_t(f) = \sum_{p,s,z,a} P_t(f|p,a)P_t(p)P_t(s|p) \times P_t(z|p,s)P(a|s,z).$$

Here, p, s, z, a are the discrete latent variables with Np, Ns, Nz, Na values respectively. At each time t, we know the F0 values indexed by p. We have to identify the underlying source playing at the p^{th} F0. Each source s has dictionaries of envelopes indexed by z. $P_t(f|p, a)$ is the fixed spectrum formed using the source-filter model as

$$P_t(f|p,a) = \frac{e_t(f|p)h(f|a)}{\sum_f e_t(f|p)h(f|a)}$$

Here, $e_t(f|p)$ consists of harmonic peaks at integral multiples of the p^{th} F0 at time t. h(f|a) is the transfer function of the a^{th} filter of a triangular mel-filter bank consisting of 20 filters uniformly distributed on the Mel-frequency scale as in [1].

3.5 Linear Discrimination Analysis (LDA) classifier:

Linear Discriminant Analysis (LDA, also known as Fisher Discriminant Analysis (FDA). LDA [6] has been widely used in face recognition, mobile robotics, object recognition and musical Instrument Classification.

In LDA, they computed a vector which best discriminates between the two classes. Linear Discriminant Analysis (LDA), searches for those vectors in the underlying space that best discriminate among classes (rather than those that best describe the data). More formally, given a number of independent features relative to which the data is described, LDA creates a linear combination of these which yields the largest mean differences between the desired classes. Mathematically speaking, for all the samples of all classes, we define two measures:

1) one is called within-class scatter matrix, as given by

$$S_w = \sum_{j\ 1}^{c} \sum_{i\ 1}^{N_j} (x_i^j - \mu_j)(x_i^j - \mu_j)^T,$$

Where x_{ji} is the i^{th} sample of class j, μ_j is the mean of class j, c is the number of classes, and N_j the number of samples in class j; and

2) the other is called between-class scatter matrix

$$S_b = \sum_{j\ 1}^{c} (\mu_j - \mu)(\mu_j - \mu)^T,$$

where μ represents the mean of all classes.

The goal is to maximize the between-class measure while minimizing the within-class measure. One way to do this is to maximize the ratio det $[S_b]$ / det $[S_w]$.

3.6 Supervised non-negative matrix factorization:

Supervised Non-Negative Matrix Factorization (S-NMF) method is one the new approach developed for the Identification/Classification of the Musical Instrument [16]. In this approach, a non-negative n × m matrix V (is considered as the features consisting of n vectors of dimension m). The non-negative n × r matrix W (basis matrix) and non-negative r × m matrix H (encoding matrix) in order to approximate the matrix V as:

$$V \approx W.H$$

Where, r is chosen such that (n + m) r < nm. To find an approximate factorization in above equation, Kullback-Leibler divergence between V and W.H is used frequently, and the optimization problem can be solved by the iterative multiplicative rules. But, the basis vectors defined by the columns of matrix W are not orthogonal. Thus, QR decomposition was utilized on W, that is W = QR, where Qn X r is an orthogonal matrix and Rr X r is an upper triangular matrix. At this time,

$$V \approx Q.H'$$

V can be written as a linear combination between an orthogonal basis and a new encoding matrix, where Q contains the orthogonal basis and H' ≈ R.H becomes the new encoding matrix. This method, however, cost a mass of computation for updating W and H iteratively and QR decomposition.

3.7 Classification Methods:

In addition to above classification techniques, some of the most important and common method for identification of Musical Instrument are also studied. DTW algorithm is powerful for measuring similarities between two series which may vary in time or speed [3]. CWT [9] too is wavelet-based feature for discrimination of various musical instrument signals. A semi-supervised learning [5] technique is also suitable for musical instrument recognition. Linear Discriminant Analysis + K-Nearest Neighbors [12] combined method has also been effectively used classification for performing automatic Musical Instrument Recognition.

4. ACKNOWLEDGMENTS

We extend our sincere thanks to Dr. S.C. Mehrotra for his helpful guidance in preparing this review.

5. REFERENCES

[1] Arora, Vipul, and Laxmidhar Behera. "Instrument identification using PLCA over stretched manifolds", Twentieth National Conference on Communications (NCC), IEEE, 2014.

[2] Bello, Juan Pablo, et al. "A tutorial on onset detection in music signals", IEEE Transactions on Speech and Audio Processing, Vol. 13.5 (2005): 1035-1047.

[3] Bhalke, D. G., C.B. Rama Rao, and D. S. Bormane. "Dynamic time warping technique for musical instrument recognition for isolated notes", International Conference on Emerging Trends in Electrical and Computer Technology (ICETECT), IEEE, 2011.

[4] Deng, Jeremiah D., Christian Simmermacher, and Stephen Cranefield. "A study on feature analysis for musical instrument classification", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 38.2 (2008): 429-438.

[5] Diment, Aleksandr, Toni Heittola, and Tuomas Virtanen. "Semi-supervised learning for musical instrument recognition", Proceedings of the 21st European Signal Processing Conference (EUSIPCO). IEEE, 2013.

[6] Eichhoff, Markus, and Claus Weihs. "Musical instrument recognition by high-level features" Challenges at the Interface of Data Analysis, Computer Science, and Optimization. Springer Berlin Heidelberg, 2012. 373-381.

[7] Eronen, Antti, and Anssi Klapuri. "Musical instrument recognition using cepstral coefficients and temporal features", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'00. Vol. 2. IEEE, 2000.

[8] Fanelli, Anna Maria, et al. "Content-based recognition of musical instruments", Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology, IEEE, 2004.

[9] Foomany, Farbod Hosseyndoust, and Karthikeyan Umapathy. "Classification of music instruments using wavelet-based time-scale features", IEEE International Conference on Multimedia and Expo Workshops (ICMEW). IEEE, 2013.

[10] Hall, Glenn Eric, Hall Hassan, and Mohammed Bahoura. "Hierarchical parametrisation and classification for musical instrument recognition", 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA). IEEE, 2012.

[11] Holzapfel, André, and Yannis Stylianou. "A statistical approach to musical genre classification using non-negative matrix factorization", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vol. 2. IEEE, 2007.

[12] Livshin, Arie, and Xavier Rodet. "Purging Musical Instrument Sample Databases Using Automatic Musical Instrument Recognition Methods", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17.5 (2009): 1046-1051.

[13] M.Goto, et al. RWC Music Database: Music Genre Database and Musical Instrument Sound Database.

[14] M.Vinyes, MTG Mass database, http://www.mtg.upf.edu/static/mass/resources.

[15] Perfecto Herrera-Boyer, Geoffroy Peeters, Shlomo Dubnov, "Automatic Classification of Musical Instrument Sounds", 2002.

[16] Rui, Rui, and Changchun Bao. "A novel supervised learning algorithm for musical instrument classification", 35th International Conference on Telecommunications and Signal Processing (TSP), IEEE, 2012.

[17] University of Iowa Musical Instrument Sample Database, http://theremin.music.uiowa.edu/index.html.