International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 756 - 761, 2014, ISSN:- 2319–8656

Automatic License Plate Recognition using OpenCV

Pratiksha Jain
Department of CSE
IGIT, GGSIPU
New Delhi, India

Neha Chopra
Department of ECE
IGIT, GGSIPU
New Delhi, India

Vaishali Gupta
Department of CSE
IGIT, GGSIPU
New Delhi, India

Abstract: Automatic License Plate Recognition system is a real time embedded system which automatically recognizes the license plate of vehicles. There are many applications ranging from complex security systems to common areas and from parking admission to urban traffic control. Automatic license plate recognition (ALPR) has complex characteristics due to diverse effects such as of light and speed. Most of the ALPR systems are built using proprietary tools like Matlab. This paper presents an alternative method of implementing ALPR systems using Free Software including Python and the Open Computer Vision Library.

Keywords: License plate, Computer Vision, Pattern Recognition, Python, OCR

1. INTRODUCTION

The scientific world is deploying research in intelligent transportation systems which have a significant impact on peoples´ lives. Automatic License Plate Recognition (ALPR) is a computer vision technology to extract the license number of vehicles from images. It is an embedded system which has numerous applications and challenges. Typical ALPR systems are implemented using proprietary technologies and hence are costly. This closed approach also prevents further research and development of the system. With the rise of free and open source technologies the computing world is lifted to new heights. People from different communities interact in a multi-cultural environment to develop solutions for mans never ending problems. One of the notable contribution of the open source community to the scientific world is Python. Intel's researches in Computer Vision bore the fruit called Open Computer Vision (OpenCV) library, which can support computer vision development.

2. PROPOSED SYSTEM

In India, basically, there are two kinds of license-plates, black characters in white plate and black characters in yellow plate. The former for private vehicles and latter for commercial, public service vehicles. The system tries to address these two categories of plates.[Reference 1]

2.1 Capture

The image of the vehicle is captured using a high resolution photographic camera. A better choice is an Infrared (IR) camera. The camera may be rolled and pitched with respect to the license plates.

Figure. 1 Example of a Number Plate with acceptable resolution

2.2. Preprocess

Preprocessing is the set algorithms applied on the image to enhance the quality. It is an important and common phase in any computer vision system. For the present system preprocessing involves two processes: Resize – The image size from the camera might be large and can drive the system slow. It is to be resized to a feasible aspect ratio. Convert Color Space – Images captured using IR or photographic cameras will be either in raw format or encoded into some multimedia standards. Normally, these images will be in RGB mode, with three channels (viz. red, green and blue).

Figure. 2 Image converted in RGB mode

After performing the steps 1 and 2, the image is passed to next component.

2.3. License Plate Extractor

This is most critical process in License Plate Recognition System. In this process we apply different techniques on image to detect and extract license plate. This process is divided in two parts.

2.3.1 License Plate Detection through Haar-like features

In image processing techniques, Haar-like features are used to recognize objects from image . If our proposed system is selected to detect only license plates then the Haar-like features are used for this purpose and no further processing is done. This technique is old and laborious and more over needs a large database to store the collected samples nearly about 10000 images of the plates and characters

2.3.2.2 License Plate Detection through Edge Detection

In the other case, if our proposed system has to recognize license plates, then the binary image is created from the image. After that following steps are performed to extract license plate from binary image:
1. Four Connected Points are searched from binary image.

2. Width/Height ratio is matched against those connected points.

3. License Plate region is extracted from image.

4. Transformation of extracted license plate is performed.

Then the extracted license plate is passed to next component for further processing.
This approach is quick and takes less execution time and memory with high a efficiency ratio. That's why we have adopted this technique in our project

Figure 3: License Plate Extraction

2.4 Character Segmentation

In this part further image processing is done on extracted license plate to remove unnecessary data. After character segmentation, the extracted license plate has only those characters that belong to license number.
This also achieved with the width height ratios matching with the contours detected on extracted number plate.

Figure 4: Character Segmentation

2.5. Optical Character Recognition

Finally, the selected blobs are send to a Optical Character Recognition (OCR) Engine, which returns the ASCII of the license number.

3.WHY OPENCV??

Advantages of OpenCV over MATLAB

- **Speed**: Matlab is built on Java, and Java is built upon C. So when you run a Matlab program, your computer is busy trying to interpret all that Matlab code. Then it turns it into Java, and then finally executes the code. OpenCV, on the other hand, is basically a library of functions written in C/C++. You are closer to directly provide machine language code to the computer to get executed. So ultimately you get more image processing done for your computers processing cycles, and not more interpreting. As a result of this, programs written in OpenCV run much faster than similar programs written in Matlab. So, conclusion? OpenCV is damn fast when it comes to speed of execution. For example, we might write a small program to detect peoples smiles in a sequence of video frames. In Matlab, we would typically get 3-4 frames analyzed per second. In OpenCV, we would get at least 30 frames per second, resulting in real-time detection.

- **Resources needed**: Due to the high level nature of Matlab, it uses a lot of your systems resources. And I mean A LOT! Matlab code requires over a gig of RAM to run through video. In comparison, typical OpenCV programs only require ~70mb of RAM to run in real-time. The difference as you can easily see is HUGE! [Reference 5].

- **Cost**: List price for the base (no toolboxes) MATLAB (commercial, single user License) is around USD 2150. OpenCV (BSD license) is *free!* Now, how do you beat that?

- **Portability**: MATLAB and OpenCV run equally well on Windows, Linux and MacOS. However, when it comes to OpenCV, any device that can run C, can, in all probability, run OpenCV.

- **Specific:** OpenCV was made for image processing. Each function and data structure was designed with the Image Processing coder in mind. Matlab, on the other hand, is quite generic. You get almost anything in the world in the form of toolboxes. All the way from financial toolboxes to highly specialized DNA toolboxes.

Despite all these amazing features, OpenCV does lose out over MATLAB on some points:

- **Ease of use:** Matlab is a relatively easy language to get to grips with. Matlab is a pretty high-level scripting language, meaning that you don't have to worry about libraries, declaring variables, memory management or other lower-level programming issues. As such, it can be very easy to throw together some code to prototype your image processing idea. Say for example I want to read in an image from file and display it**.**

- **Memory Management:** OpenCV is based on C. As such, every time you allocate a chunk of memory you will have to release it again. If you have a loop in your code where you allocate a chunk of memory in that loop and forget release it afterwards, you will get what is called a "leak". This is where the program will use a growing amount of memory until it crashes from no remaining memory. Due to the high-level nature of
Matlab, it is "smart" enough to automatically allocate and release memory in the background.
- -Matlabs memory management is pretty good. Unless your careful with your OpenCV memory allocation and releasing, you can still be frustrated beyond belief.

- **Development Environment**: Matlab comes with its own development environment. For OpenCV, there is no particular IDE that you have to use. Instead, you have a choice of any C programming IDE depending on whether you are using Windows, Linux, or OS X. For Windows, Microsoft Visual Studio or NetBeans is the typical IDE used for OpenCV. In Linux, its Eclipse or NetBeans, and in OSX, we use Apple's Xcode**.**

- **Debugging**:Many of the standard dedugging operations can be used with both Matlab and OpenCV: breakpoints can be added to code, the execution of lines can be stepped through, variable values can be viewed during code execution etc. Matlab however, offers a number of additional debugging options over OpenCV. One great feature is that if you need to quickly see the output of a line of code, the semi-colon at the end can be omitted. Also, as Matlab is a scripting language, when execution is stopped at a particular line, the user can type and execute their own lines of code on the fly and view the resulting output without having to recompile and link again. Added to this is are Matlab's powerful functions for displaying data and images, resulting in Matlab being our choice for the easiest development environment for debugging code.

3.1 Conclusion

	Matlab	OpenCV
Ease of Use	9	3
Speed	2	9
Resources Needed	4	9
Cost	4	10
Development Environment	8	6
Memory Management	9	4
Portability	3	8
Development of usefull programming skills	3	8
Help and Sample Code	8	9
Debugging	9	5
Total:	59	71

Figure 5: NECKBEARD INDEX SCORES

From the final scores we can see that OpenCV has the edge over Matlab for image and video processing development . Although Matlab has an easy learning curve, built in memory management, a great help section, it is very slow to execute code, and is expensive to get started in. While OpenCV can be difficult to debug and requires much "housework code" needed for memory management, header files, etc., it wins out due to its free cost, the magnitude of sample code available on the internet, the short development path from prototype code to embedding code, the useful programming skills learnt from its use, and its super-fast speed. Matlab is a more "generic" programming language in that it was designed for many uses, demonstrated by its numerous toolboxes ranging from financial to specialized DNA analyzing tools. On the other hand, OpenCV was made for image processing. Each function and data structure was designed with the image processing coder in mind**.**

4. PROPOSED SOLUTION

4.1 Assumptions

The objective of our thesis is to detect and recognize license plate. Our application is based on following assumptions:

1. Maximum expected distance between car and camera: 5 meters.

2.Minimum Camera angle: 90 degree (looking straight at the license plate).

3.Image should be captured in daylight.

4.Minimum Camera resolution: 3 Mega Pixel.
It is expected that it would not work efficiently during night time, rainy and cloudy days because mobiles cameras are not equipped with proper lightning. It is also expected that it will give results with decreasing accuracy with angles deviating significantly from the 90-degree (ideal) angle.

5. The new algorithm proposed for character recognition would give results with considerable percentage of errors on implementation.

6. The efficiency of the proposed system can be measured only in terms of number of license plates successfully and correctly recognized which can only be measured upon implementation.

7. Efficiency and Performance of new system may decline due to discard of OCR library but the memory requirements will decrease and also the effort for installing, configuring and running the system would decrease.

4.2 New Components of Proposed System as compared to traditional system

- **DETECTION ALGORITHM**
  We are designing this system specifically for the new proposed high security number plates which have black boundary across the number plate and also have a uniform font all across the country. So we are going to utilize this black boundary in our system by using edge based license plate detection method in our system. traditionally haar like features are used for detection. This algorithm needs a large number of license plate images which are manually obtained from a number of images including the backgrounds. It requires a larger memory to run, which is not suitable for embedded systems. Another problem with the systems using AdaBoost is that they are slower than the edge-based methods. This system is very sensitive to the distance between the camera and the license plate as well as the view angle. So we can eliminate all the above problems by using edge based detection method for our system. however detection rate of edge based method is slightly less than haar like features. This can be supported by the study conducted by some research students of Linnaeus university. Haar like feature were 96% accurate while edge based method was 87% accurate.

- **OCR LIBRARY NOT USED**
  In traditional system OCR Library is used which has to be installed, configured and run and which actually recognize the characters. We are not using this library. Instead we are developing our own algorithm for character reading. also OCR engines occupy more than 25 MB space and configuration of OCR engine has to be done with the source code. Compiler takes quite long time in compilation of typical OCR code because of the specific quality checks, spell checks, valid word checks etc. these checks are not required in ALPR case because spell checks, valid word checks are useless in case of number plates. so our algorithm is simple, fast and occupies less memory than an OCR engine. also it is expected that it will provide correct results upon implementation

4.3 Proposed Algorithm

DESCRIPTION OF THE NEW ALGORITHM FOR CHARACTER RECOGONITION

In this part, character segmented license plate is passed to optical character recognition algorithm designed by us which uses a matrix transformation of the pixel value of the binary and thus applying various filtration and extraction techniques which uniquely identifies the characters. OCR Algorithm returns license plate in text format. Which is later stored in a text file thus reducing the space in the memory storage.[Reference 3]

- Our algorithm uses a 3-4 MB database of 36 files(images).
- These 36 images are samples containing capital alphabets(A-Z) and numerals(0-9).
- These images will be colored images but only of one color say red. So pixel values where there is character is 255,0,0.
- and where the space is empty the value is 255,255,255. then the characters obtained after character segmentation are mapped with the characters in the data base one by one
- The character obtained from segmentation is mapped to a matrix one by one.
- then this matrix is compared with the sample images in database one by one.
- if the character matches then the value of the character is returned. Else next character is matched.
- if any of the 36 characters don't match with the image then either there is a distorted image or the number plate is invalid. In this condition a message will be returned.
- The matrix used will be preferably 20x20.
- for mapping between sample image and actual character we are using green intensity pixels. Because their value is 0 at every point where there is character and 255 where there is white background.
- we could have used blue intensity as well.
- this algorithm will thus possibly be able to detect similar characters like 8 and B because percentage of matching of one character will be higher than other.
- It is assumed that if any image is matched with 70-80% pixel intensities we assume that character matches
- then matrix is refreshed and new character gets copied in matrix.

the process continues until all the characters in license plate gets matched.

4.3.1 Algorithm for OCR Reader

```
OCRReader(image temp)
{
        Int mat[30][30]
        for(y=0, y<temp->height , y++)
        {
                for (x=0 , x<temp->width , x++)
                {
                /value is a structure/
                value=get RGB values of temp
                /b,g,r are integers/
                b=value.val[0]
                g=value.val[1]
                r=value.val[2]

                mat[y][x]=g;
                }
        }
        stringcopy(file,"folder of 36 files")
        for(int j=0 , j<36 , j++)
        {
                count=0
                stringcopy(file,"folder of 36 files")
                ext=get file j
                l= length of string (file)
                file[l]=ext
                file[l+1]='\0'
                file=file+"jpg"
         lchar=create image of frame of given size
         lchar= load image of name  file +"jpg"
                for(y=0; y<lchar->height; y++)
                {
                for (x=0;x<lchar->width;x++)
                {
                value= get RGB values of lchar
                b=value.val[0]
                g=value.val[1]
                r=value.val[2]

                l_mat[y][x]=g;
                }
                }

        for(y=0;y<30;y++)
        {
                for(x=0;x<30;x++)
                {
                        if(mat[y][x]==l_mat[y][x])
                        count++;
                }
        }
        if(count>400)
                {
                cout<<ext<<"in";
                }
      }
}
```

4.4 Asymptotic analysis of Algorithm

Complexity of above code is O(mn^2). Where m=36(A-Z , 0-9)and n is the pixel resolution. this is same as complexity of OCR reader.
But In traditional System OCR engine has database of 2^16 symbols(Unicode). So there value of m=2^16. Hence significant reduction in Time complexity. also since database is of 36 symbols instead of 2^16 it results in significant reduction in Space complexity.

5. CONCLUSION

The message of this research is to show that free and open source technologies are matured enough for scientific computing domains**.** The system works satisfactorily for wide variations in illumination conditions and different types of number plates commonly found in India. It is definitely a better alternative to the existing proprietary systems, even though there are known restrictions

5.1 Future Work

Currently We have proposed the algorithms for our ALPR system. In future we would implement this system on Open CV library and would also do the performance check of the system designed. We would do the performance analysis in terms of number of plates successfully recognized. So far the algorithms looks good and suitable but if the OCR algorithm won't work than we will try to give some new algorithm or would do the comparative study of different OCR present in the market and would try to choose the best among them and implement the system.

6. ACKNOWLEDGMENTS

Our Sincere Thanks to Dr Kalpana Yadav ,our mentor for providing her guidance and cooperation in this Research.

7. REFERENCES

[1] A. Conci, J. E. R. de Carvalho, T. W. Rauber, " A Complete System for Vehicle Plate Localization, Segmentation and Recognition in Real Life Scene" , IEEE LATIN AMERICA TRANSACTIONS, VOL. 7, NO. 5,SEPTEMBER 2009

[2] Ahmed Gull Liaqat, "Real Time Mobile License Plate Recognition System" IEEE White paper California, VOL.2 2011-12-05,Linnaeus University.

[3] Ondrej Martinsky (2007). "Algorithmic and mathematical principles of automatic number plate recognition systems" (PDF). Brno University of Technology. http://javaanpr.sourceforge.net/anpr.pdf.

[4] P. Kreling, M. Hatsonn "A License Plate Recognition algorithm for Intelligent Transportation System applications". University of the Aegean and National Technical University of Athens. 2006. Archived from the original on 2008-04-20.

[5] K.M Sajjad , "ALPR Using Python and Open CV" Dept Of CSE, M.E.S College of Engineering Kuttipuram, kerala.2008-06-21

[6] Nicole Ketelaars "Final Project : ALPR", 2007-12-11

[7] Steven Zhiying Zhou , Syed Omer Gilani and Stefan Winkler "Open Sourc framework Using Mobile Devices" Interactive Multimedia Lab, Department of Electrical and Computer Engineering National University of Singapore, 10 Kent Ridge Crescent, Singapore 117576

[8] Yunggang Zhang, Changshui Zhang "A new Algorithm for Character Segmentation Of license plate" Beijing University,China, 2007-5-8

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 762 - 768, 2014, ISSN:- 2319–8656

Analysis the Effect of Educational Package on Promotion of Protective Behaviors in Exposure to Dust Phenomenon by SPSS Software

Ali Ramezankhani[1]
Department of Public Health, Faculty of Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Kobra Doostifar[2*]
Department of Public Health, Shushtar Faculty of Medical Sciences, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

Saeed Motesaddi Zarandi[3]
Department of Environmental Health, Faculty of Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Tayebeh Marashi[4]
Department of Public Health, Faculty of Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Nezhat Shakeri[5]
Department of Biostatistics, Faculty of Paramedicin, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Maryam Parsanahad[6]
Department of nutrition, Shushtar Faculty of Medical Sciences, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

* Corresponding Author: Kobra Doostifar, Department of Public Health, Shushtar Faculty of Medical Sciences, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran.

ABSTRACT

Background: dust phenomenon, especially in the summer, is a serious problem in Khuzestan province and has adverse effects on health, environment and economic. Behavior change is the base for health associated risks prevention and one of the models for behavior change in individual level is Health Belief Model. The aim of this study was to analyze the effect of educational package on promotion of protective behaviors in exposure to dust phenomenon in Ahvaz teachers by SPSS software.

Methods: This was an experimental study in which 200 teachers randomly were divided into two groups, case and control groups [n=100, in each group]. Data were collected by a Health Belief Model questionnaire whose validity and reliability were confirmed by content validity method and Cronbach's Alpha test. Before the educational intervention, questionnaire was completed by two groups and educational requirements of subjects were detected and an educational package was designed and implemented for 4 week. The control group received no intervention. After a month the effect of educational package on study variables was evaluated. Data were analyzed with SPSS statistical software version 17, by descriptive and analytical tests.

Result: Mean age of case and control groups were 39.75±6.95 and 39.78±7.02 years, respectively. There was no significant association between marriage and behavior, but there was a significant association between employment number of years and behavior [p=0.03], education and behavior [p=0.03]. Based on the findings of this study there was a significant association between the knowledge, health belief model components and behavior of the study subjects, before and after the intervention [p< 0.001].

Conclusion: designing and implementation of an educational package based on health belief model can promote the knowledge and protective behaviors in exposure to dust particles effectively.

Keywords: education, educational package, protective behaviors, dust phenomenon

1. INTRODUCTION

Scientific researches in the past two decades have shown that particles are one of the specific pollutants [1]. Results of a study by the World Health Organization in Berlin, Copenhagen and Rome showed that particles smaller than 2.5 microns in diameter, seriously affect the health and increase death due to respiratory disease, cardiovascular disease and lung cancer [2]. According to the World Health Organization report over 2 million people developing premature death in every year [3]. Most important effects of dust are allergy of eyes, nose and throat, respiratory tract infections, headache, nausea, allergic reactions, chronic respiratory disease, lung cancer, heart disease and damage to other organs of the body. Dust in the long-term can change the mood. Aggression and depression are also other effects of dust [4]. Annette peters showed an association between heart disease and air particles based on epidemiological evidences. This study showed that daily changes in air particles concentration are closely associated with cardiovascular related deaths, hospital admit, cardiovascular disease symptom exacerbation [5]. Chinese researchers in 2007 investigated the impact of particles smaller than 2.5 microns in samples that collected from Asian dust on macrophage and lung cells DNA of mice.

The results showed that extract of these particles leaded to DNA damage in these cells [6]. In a study on students of 850 schools in United Arab Emirates the prevalence of asthma and allergy were 13.6% and 73%, respectively and there was a significant association between dust and mentioned diseases [7]. Some recent epidemiological studies suggested that long-term transport of dust particles is associated with increased daily mortality in Seoul [8] and Taipei and Taiwan [9] and respiratory and cardiovascular diseases [8]. although , the dust particles contribute to drinking water contamination and therefore gastrointestinal disease [10].

Dust occurrence increases some heavy metals such as lead concentrations up to 3-fold [11]. Also the concentration of toxic metals, mercury and arsenic, greatly increased [12]. Air -borne microorganisms concentration in dusty days increase and most of these microorganisms are pathogen and cause disease.

Iran is located in an area with a dry climate and over 30% of the country is arid and semi-arid area [14]. In the past few years the country has been exposed to the dust phenomenon. This country because of neighboring with the wide expanse of desert is adversely affected by this phenomenon. One of the areas that has been affected by this phenomenon is Khuzestan province that is located in southwestern of Iran [15]. This phenomenon has been associated with some problems and rise in adverse effects on health, environment and economic [15]. In the dusty days, admit of patients with pulmonary disease to health centers in Ahvaz has showed 70 percent increase [1]; One way for reducing the incidence of diseases caused by dust is educational interventions. Health education experts use the appropriate models in order to health education interventions design, one of this models is Health Belief Model. The aim of this study was design and implementation of educational package based on health belief model and evaluation of its effects on protective behaviors in teachers by SPSS software. In this study educational package was an educational program that has been designed based on educational needs of subjects in order to prepare the subjects for implementation of protective behaviors in exposure to dust phenomenon.

2. METHODS

This was an experimental and analytical study [before and after] that has been conducted in Ahvaz. Two hundred teachers randomly were divided into two groups, case and control groups [n=100, in each group]. The inclusion criteria included: employment for at least three years, lack of respiratory disease and cardiovascular disease and satisfaction for participation in the study. Exclusion criteria included: unsatisfaction for participation in the study and nonparticipation in the educational sessions. Data were collected by a questionnaire that was designed according to the health belief model constructs. The questionnaire contained 78 questions in four parts. This parts included questions regarding to individual characteristics [19 questions], knowledge [14 questions, score range=14-32], health belief model constructs [34 questions] and protective behaviors in exposure to dust phenomenon [11 questions, score range=11-55], respectively. In part 3, questions included: perceived susceptibility, perceived severity and perceived benefits each with 7 questions [score range = 7-35], perceived barriers with 6 questions [score range = 6-36], cue to action with 2 questions [score range = 2-10] and self-efficacy with 4 questions [score range = 4-20].

The validity of questionnaire was evaluated by means of face validity and content validity methods. Face validity was evaluated by means of relevance, simplicity and clarity of questions. Questionnaire was evaluated by 10 experts [included 5 experts in health education, 5 experts in environmental health, 1 epidemiologist and 2 experts in Biostatistics]. Content validity was evaluated by means of Lawshe' s technique. Questionnaire reliability was evaluated by means of Cronbach's Alpha test in 20 teachers that were same with study population in demographic characteristics.

Cronbach's Alpha coefficient of questionnaire parts was detected: Knowledge: 0.76, perceived susceptibility: 0.73, perceived severity: 0.88, perceived benefits: 0.72, perceived barriers: 0.77, cue to action: 0.71, perceived Self-efficacy: 0.71 and protective behaviors: 0.71.

The questionnaire was used before and after the educational package implementation to determine the perceived knowledge, sensitivity, severity, benefits, barriers and self-efficacy, and behavior of subjects. Data were collected by a questionnaire in interview method before intervention in case and control group. Then data were analyzed and the educational needs of subjects were detected and educational package was designed. Educational package included an educational booklet, pamphlet and CD that represented essential information in relation to dust particles, disease prevention and protective behaviors. Then researcher represent educational package to the case group in four sessions [each sessions was 90 minutes]. Educational methods were the lecture, questioning and responding and showing the video clip. Immediately and two months after the educational intervention, subjects data were collected by questionnaire and were analyzed. The control group received no intervention. Data were analyzed with SPSS statistical software version 17, by frequency distribution, correlation coefficient, t-Student, Chi-square, Mann-Whitney and Repeated measures tests.

3. RESULTS

Two hundred teachers were participated in this study. Mean age of case and control groups were 39.75±6.95 and 39.78±7.02 years; respectively. Age 40-49 years had the most frequency in the case group [46%] and the control group [45%]. In the two groups more present of subjects were married [82% in the case group and 81% in the control group]. In the two groups most of subjects had Bachelor's degree. More present of subjects had two children [47% in the case group and 46.3% in the control group] and less present of subjects had four children. Most subjects didn't receive education about dust phenomenon and protective behaviors. Age, marriage, education,

Number of children and previous education about dust phenomenon were not significantly difference between cases and controls.

There was no significant association between marriage and behavior but there was a significant association between employment number of years and behavior (p=0.03) and also between education and behavior (p=0.03). In the two groups the most used sources for information about protective behaviors in the exposure to dust particles were radio and television and there was no significant difference between two groups table 1.

Table 1. Information sources regarding the dust phenomenon in teachers, Ahvaz

Groups / Information source	cases		controls		p-value
	yes	no	yes	no	
Radio & television	93	7	95	5	0.552
Newspaper & magazine	43	57	40	60	0.667
family	58	42	58	42	1
coworkers	57	43	58	42	0.886
friends	55	45	57	43	0.776
Book & booklet	32	68	31	69	0.115
Physician and staff of health center	36	64	35	65	0.077
internet	47	53	45	55	0.777

The mean of knowledge, perceived susceptibility, perceived severity, perceived benefits, perceived barriers, perceived self-efficacy, cue to action and behavior score were not significantly different between cases and controls before intervention. Whereas, immediately and two months after the educational intervention there was a significant difference between cases and controls in mentioned variables [p=0.001] [table 2, 3].

Before the intervention, 16% of cases often didn't leave the home in dusty days but after the intervention 57% of cases often didn't leave the home in dusty days. Before the intervention 70% of cases sometimes educated their students in relation to air pollution but after the intervention 75% of cases often educated their students. Before the intervention, only 2% of cases have been eaten more amount of fruit and vegetable in dusty days but after the intervention the rate increased to 41%. Before the intervention, only 3% of cases have been eaten more amount of milk in dusty days.

4. DISCUSSION

One of the most important air pollutants are dust particles and high concentrations of particles in dust storms causing sinusitis, bronchitis, asthma, allergy and damage to the defensive function of macrophages, thereby leading to an increase in hospital infections [19]. The purpose of the present study was implementation of protective behaviors when dust phenomenon occurs. To the best of our knowledge the effect of this educational method on protective behaviors in exposure to dust particles has not been investigated in previous studies.

Before the intervention, protective behaviors of teachers in exposure to dust phenomenon were in intermediate level. But significant difference between behavior score of cases and controls after intervention showed the positive effect of educational package on promotion of protective behaviors in case group. In Araban et al. study after the intervention behavior score was significantly different between case and control groups [20]. The results of Giles et al. meeting in Canada on strategies for reducing the adverse effects of air pollution on health, entitled "The decision to effective intervention", showed that personal behavior modification and pollutants exposure reduction are appropriate approaches for reducing the adverse effects of air pollution [21]. Sexton study showed that on dusty days persons changed their behavior by reducing time spent outdoors by 18% or 21 minutes [22].

In the present study before the intervention the two group's sources of information about protective behaviors in the exposure to dust particles were radio, television and family. Significant difference between knowledge score of two groups after the educational intervention was due to the educational sessions about protective behaviors in exposure to dust phenomenon and this educational sessions promoted the knowledge of case group about protective behaviors. These results are in line with the use of Health Belief Model in researches about diabetes control and self-care and promotion of knowledge after the educational intervention [23, 24]. Boonkuson et al. showed that protective behaviors in exposure to health problems depends on the knowledge and attitude [25]. Pazira et al. reported that a part of Tehran population knowledge about air pollution and protective behaviors was in low level [26].

In the health belief model constructs the perceived susceptibility score before intervention was the same in both groups. After intervention perceived susceptibility score was significantly different between case and control groups [p=0.001]. This finding is consistent with increased perceived susceptibility in researches about the osteoporosis prevention [27] and diet care [24].

Also, perceived severity score before intervention in two groups showed that teacher's perception from seriousness of illnesses caused by dust particles was over the average, probably due to the illness of friends or coworkers or damages caused by dust particles. The dramatic increase in the perceived severity score of the case group seems to be due to the teachers' participation in educational sessions

and providing educational package included showing video clip, booklet and pamphlet, mention to importance of protective behaviors on the dusty days, high cost of pulmonary, cardiovascular and gastrointestinal tract diseases. In the other studies perceived severity has been increased similarly [23, 27]. Also in the Praphant et al. study perceived severity was in moderate level [60.6%] [28].

Table 2. Comparing knowledge and behavior scores regarding the protective behaviors in exposure to dust phenomenon in teachers, Ahvaz

variable	group	Before intervention mean± SD	Immediately after intervention mean±SD	2 month after intervention mean±SD	Repeated Measures test
knowledge	Case	53.81 ±3.43	58/77 ±1/44	58.13 ± 2.54	P <0.001
	Control	53.65 ±3.5	53/15 ±2/66	53.48 ± 3.64	P< 0.2
	Independent sample t test	P <0.745	P <0/001	P <0.001	
behavior	Case	33 ±4.14	37/81 ±3/77	38.98 ± 2.97	P <0.001
	Control	34.13 ±4.7	34/99 ±4/08	34.22 ± 4.66	P <0.176
	Independent sample t test	P <0.073	P <0/001	P <0.001	

Table 3. Comparing health belief model constructs scores regarding the protective behaviors in exposure to dust phenomenon in teachers, Ahvaz

variable	group	Before intervention mean± SD	Immediately after intervention mean±SD	2 month after intervention mean±SD	Repeated Measures test
Perceived susceptibility	Case	27.02 ±2.58	29.6 ±2.29	29.6 ± 2.28	P <0.001
	Control	27.38 ±2.74	27.35 ±2 .74	27.41 ± 2.78	P <0.988
	Independent sample t test	P <0.341	P <0.001	P <0.001	
Perceived severity	Case	28.75 ±1.97	31.7 ±2.43	31.46 ± 2.23	P <0.001
	Control	29.03 ±2.4	28.84 ±2.47	29.04 ± 2.44	P <0.792
	Independent sample t test	P <0.369	P <0.001	P <0.001	
Perceived benefits	Case	28.09 ±2.87	30.54 ±3.02	30/39 ± 2/95	P <0.001
	Control	28.75 ±2.91	28.71 ±2.94	28/66 ± 2/89	P <0.978
	Independent sample t test	P <0.109	P <0.001	P <0.001	
Perceived barriers	Case	16 ±4.51	18.12 ±4.73	18.19 ± 4.51	P <0.002
	Control	16.37 ±4.62	16.36 ±4.57	16.27 ± 4.55	P <0.943
	Independent	P <0.557	P <0.011	P <0.006	

	sample t test				
Perceived self-efficacy	Case Control Independent sample t test	12.51 ±2.15 13.07 ±2.49 P <0.091	15.19 ±2.2 13.02 ±2.57 P <0 .001	15.07 ±2 .27 12.98 ± 2.52 P <0.001	P <0.001 P <0.97
Cue to action	Case Control Independent sample t test	7 ±1.22 7.06 ±1.48 P <0.755	7.4 ±1.1 7.06 ±1.48 P < 0.041	7.45 ± 1.11 7.04 ± 1.32 P <0.049	P <0.007 P <0.13

The results showed that before the intervention teachers' perceived benefits of protective behaviors in dusty days in both groups were in good condition but after the intervention perceived benefits score has increased in the case group. Because the protective behaviors in the exposure to dust phenomenon are not time consuming and expensive and are simple and don't need the physician visit, can be useful in promotion of perceived benefits. Araban et al. showed that perceived benefits increased by improvement in stage of change [20]. Qaderi et al. reported that perceived benefits score increased in the case group after the intervention [29].

Perceived barriers of protective behaviors in both case and control groups were moderate before intervention; but there were significant differences between the perceived barriers of the two groups after the intervention due to the effect of the education. Most of the teachers' perceived barriers for protective behaviors in exposure to dust phenomenon included the unavailability of respiratory mask, discomfort and shortness of breath and nose sweating because of the mask, financial difficulties in buying more fruit and vegetable in a dusty days, financial difficulties due to stay at home and problems related to communication with coworkers. in Araban et al. study two barriers , delay in doing things and need to enter the crowded areas of city, After education in the intervention group changed, while these barriers in the control group was not significant [20]. Koch showed that elimination of perceived barriers increased walking in diabetic patients [30].

Other Health Belief Model construct was perceived self-efficacy. Self-efficacy is the beliefs of person about their abilities to control events that affects their life [31]. Teacher's self-efficacy score was in low level in case and control groups before intervention. But Teacher's self-efficacy score was significantly different between cases and controls after intervention that this was due to the effect of education on self-efficacy and promotion of protective behaviors in the case group. Araban et al. reported that self-efficacy was significantly higher in case group after the intervention [20].

Education based on Health Belief Model promoted the teachers protective behaviors in exposure to dust phenomenon by promoting the perceived susceptibility, severity, benefits and barriers using a variety of educational methods and educational package. On the other hand, Stimuli or cues to action encouraged teachers to the protective behaviors. Also, present study showed that the media played an important role in attracting teachers to protective behaviors. Drakshyani et al. in their study on schools and colleges teachers in India showed the Necessity of public health education programs through the mass media [32]. Khorsandi et al. reported that radio and television programs are the most important cues to action in reducing the risk of osteoporosis [33] .The present study designed an educational package in order to promote the teachers behaviors but similar research should be conducted in other parts of the country.

5. CONCLUSION

The findings of this study showed that the designed educational package was effective in promoting the knowledge and protective behaviors in teachers. Therefore, health behavior education in other people, especially in high-risk groups is important to maintenance of protective behaviors in exposure to the dust phenomenon.

6. ACKNOWLEDGEMENT

The source of data used in this paper was from MSc thesis. The authors express sincere thanks to the teachers because of participation and cooperation in the study.

11. REFERENCES

1. Colls J. Air pollution.2nd ed Taylor, Francis, Inc, London and New York. 2003.p.4.

2. World Health Organization. Particulate matter air pollution: how it harms health. 2005 April. Fact sheet EURO/04/05. Available from: http: //www.euro.who.int /document/ mediacentre/ fs0405e.pdf. Accessed July 16, 2013.

3. Department of Public Health and Environment. World Health Organization Geneva Switzerland. Urban outdoor air pollution database. 2012. Available from: http: // www. who. int/ phe 2012. Accessed Jul 30, 2012.

4. Griffin DW, Kellogg CA. Dust storms and their impact on ocean and human health: dust in Earth's atmosphere. EcoHealth. 2004;1[3]:284-95.

5. Peters A. Particulate matter and heart disease: evidence from epidemiological studies. Toxicology and applied pharmacology. 2005;207[2]:477-82.

6. Meng Z, Zhang Q. Damage effects of dust storm PM< sub> 2.5</sub> on DNA in alveolar macrophages and lung cells of rats. Food and chemical toxicology. 2007;45[8]:1368-74.

7. Bener A, Abdulrazzaq Y, Al-Mutawwa J, Debuse P. Genetic and environmental factors associated with asthma. Human biology. 1996:405-14.

8. Kwon H-J, Cho S-H, Chun Y, Lagarde F, Pershagen G. Effects of the Asian dust events on daily mortality in Seoul, Korea. Environmental Research.2002;90[1]:1-5.

9. Ichinose T, Yoshida S, Hiyoshi K, Sadakane K, Takano H, Nishikawa M, et al. The effects of microbial materials adhered to Asian sand dust on allergic lung inflammation. Archives of environmental contamination and toxicology. 2008; 55[3]: 348-57.

10. Kellogg CA, Griffin DW, Garrison VH, Peak KK, Royall N, Smith RR, et al. Characterization of aerosolized bacteria and fungi from desert dust events in Mali, West Africa. Aerobiologia. 2004;20[2]:99-110.

11. Viana M, Kuhlbusch T, Querol X, Alastuey A, Harrison R, Hopke P, et al. Source apportionment of particulate matter in Europe: a review of methods and results. Journal of Aerosol Science. 2008;39[10]:827-49.

12. Wang Y, Zhang X, Arimoto R, Cao J, Shen Z. Characteristics of carbonate content and carbon and oxygen isotopic composition of northern China soil and dust aerosol and its application to tracing dust sources. Atmospheric Environment. 2005;39[14]:2631-42.

13. Schlesinger P, Mamane Y, Grishkan I. Transport of microorganisms to Israel during Saharan dust events. Aerobiologia. 2006;22[4]:259-73.

14. Modarres R. Regional maximum wind speed frequency analysis for the arid and semi-arid regions of Iran. Journal of Arid Environments. 2008;72[7]:1329-42.[In persian]

15. Zarasvandi A, Moore F, Nazarpour A. Mineralogy and morphology of dust storms particles in Khuzestan province: XRD and SEM analysis concerning. Iranian journal of crystallography and mineralogy. 2011;19[3]:511- 8.[In persian]

16. Hossein Gholizadeh N. A effect of intervention based on HBM on improving of knowledge, attitude and practice among students in Tehran [dissertation]. School of public health: Tehran university of medical sciences, 2010.[In persian]

17. Mirzaei E. Health education and health promotion in textbook of public health. Tehran: Rakhshan. 2004.[In persian]

18. Taheri Aziz M. Effectiveness of Designed Health Education Package on Healthy Behaviors of Patients with Tuberculosis at Pasteur Institute of Iran [dissertation]. Tehran: Tarbiat modares of Medical sciences; 2004. p.67-8.[In persian]

19. Al-Hurban AE, Al-Ostad AN. Textural characteristics of dust fallout and potential effect on public health in Kuwait City and suburbs. Environmental Earth Sciences. 2010;60[1]:169-81.

20. Araban M. Design and Evaluation of a Theory-Based Educational Intervention on Behavioral Improvement in Pregnant Women in Terms of Exposure to Air Pollution [Dissertation]. Tehran: Tarbiat Modares University, Faculty of Medical Sciences; 2013. [Text in Persian]

21. Giles LV, Barn P, Kunzli N, Romieu I, Mittleman MA, van Eeden S, et al. From good intentions to proven interventions: effectiveness of actions to reduce the health impacts of air pollution. Environmental health perspectives. 2011; 119[1]:29.

22. Sexton AL. Responses to Air Quality Alerts: Do Americans Spend Less Time Outdoors? [Dissertation]. Minnesota: Department of Applied Economics, University of Minnesota; 2011.

23. Mohebi S, Sharifirad G, Hazaveyee S. The effect of educational program based on Health Belief Model on diabetic foot care. Int J Diab Dev Ctries. 2007; 27:18-23.[In persian]

24. Kamrani A. The effect of educational diet on nutrition type2 diabetes based on Health Belief Model [Dissertation]. Faculty of Public Health, Isfahan University of Medical Science ,2006.[In persian]

25. Boonkuson T. Comparisons of behavior on protection of health problems caused by rock dust of the population with difference on personal factors and social and economic factors in the rock crusher plants, Saraburi province. [dissertation]. Project joint research of nursing college attached to institute of development of public health personnel, 1994.

26. Pazira M,Ghanbari R, Askari E. Survey of knowledge, attitude and practice about air pollution among of people lives in Tehran and some activity of emergency. Conference on air pollution and effects on health. 2005 Feb:1- 2; Tehran, Iran.

27. Saeedi M. The survey of educational program based on health belief model on preventive osteoporosis [Dissertation]. School of Public Health: Isfahan University of Medical Science, 2005.[In persian]

28. Praphant A. Preventive behaviors form dust among workers in lime factories and stone crushing mills, Nakhon Si Thammarat province. [dissertation]. College of Public Health: Chulalongkorn University. 2003.

29. Amal KA, Dalal MAR, Ibrahim KL. Effect of educational film on the health belief model and self-examination practice. East Mediterr Health J. 1997;3[3]:435-44.

30. Koch J. The role of exercise in the African-American woman with type 2 diabetes mellitus: application of the health belief model. J Am Acad Nurse Pract 2002;14[3]:126–9.

31. Kazdin AE. Encyclopedia of Psychology. New York: Oxford University Press; 2000. p. 212–3.

32. Drakshyani Devi K, Venkata Ramaiah P. Teacher's knowledge and practice of breast self examination. Indian J Med Sci 1994;48[12]:284–7.

33. Khorsandi M, Shamsi M, Jahani F. The Survey of Practice About Prevention of Osteoporosis Based on Health Belief Model in Pregnant Women in Arak City. Journal of Rafsanjan University of Medical Sciences. 2013;12[1]:35-46.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 769 - 773, 2014, ISSN:- 2319–8656

Illumination Invariant Face Recognition System using Local Directional Pattern and Principal Component Analysis

Latha B
Department of Computer Science and Engineering
Roever College of Engineering and Technology,
Perambalur Tamilnadu
India – 621 220.

Dr. Punidha R
Department of Computer Science and Engineering
Roever College of Engineering and Technology,
Perambalur Tamilnadu
India – 621 220.

Abstract: In this paper, we propose an illumination-robust face recognition system using local directional pattern images. Usually, local pattern descriptors including local binary pattern and local directional pattern have been used in the field of the face recognition and facial expression recognition, since local pattern descriptors have important properties to be robust against the illumination changes and computational simplicity. Thus, this paper represents the face recognition approach that employs the local directional pattern descriptor and two-dimensional principal analysis algorithms to achieve enhanced recognition accuracy. In particular, we propose a novel methodology that utilizes the transformed image obtained from local directional pattern descriptor as the direct input image of two-dimensional principal analysis algorithms, unlike that most of previous works employed the local pattern descriptors to acquire the histogram features. The performance evaluation of proposed system was performed using well-known approaches such as principal component analysis and Gabor-wavelets based on local binary pattern, and publicly available databases including the Yale B database and the CMU-PIE database were employed.

Keywords: Face Recognition; Local Directional Pattern; Principal Component Analysis.

1. INTRODUCTION

Face recognition has become one of the most popular research areas in the fields of image processing, pattern recognition, computer vision, and machine learning, because it spans numerous applications [1, 2]. It has many applications such as biometrics systems, access control systems, surveillance systems, security systems, credit-card verification systems, and content-based video retrieval systems. Up to now, main algorithms have been applied to describe the faces: principal component analysis (PCA) [3], linear discriminant analysis (LDA) [4], independent component analysis (ICA) [5] and so on. Generally, face recognition systems can achieve good performance under controlled environments. However, face recognition systems tend to suffer when variations in different factors such as varying illuminations, poses, expression are present, and occlusion. In particular, illumination variation that occurs on face images drastically degrades the recognition accuracy To overcome the problem caused by illumination variation, various approaches have been introduced, such as preprocessing and illumination normalization techniques [6], illumination invariant feature extraction techniques [7], and 3D face modeling techniques [8]. Among abovementioned approaches, local binary pattern (LBP) [9] has received increasing interest for face representation in general [10]. The LBP is a non-parametric kernel which summarizes the local spatial structure of an image. Moreover, it has important properties to be tolerant against the monotonic illumination changes and computational simplicity. More recently, the local directional pattern (LDP) method was introduced by Jabid et. al for a more robust facial representation [11]. Because LBP is sensitive to non-monotonic illumination variation and also shows poor performance in the presence of random noise, they proposed the LDP descriptor as face representation and also demonstrated better performance compared to LBP.

In this paper, we present a novel approach for achieving the illumination invariant face recognition via LDP image. Most of previous face recognition researches based on LBP utilized the descriptor for histogram feature extraction of the face image. Similar to LBP, LDP descriptor is also utilized to extract the histogram facial features in previous researches [11]. However, this paper uses the LDP image as a direct input image of 2D-PCA algorithms for illumination-robust face recognition system. The proposed approach has an advantage that the illumination effects can be degraded by using binary pattern descriptor and 2D-PCA is more robust against illumination variation than global features such as PCA and LDA since 2D-PCA is a line-based local feature. The performance evaluation of the proposed system was carried out using the Yale B database [12] and the CMU-PIE illumination/light database [13]. Consequently, we will demonstrate the effectiveness of the proposed approach by comparing our experimental results to those obtained with other approaches.

2. PROPOSED APPROACH

This paper aimed to improve face recognition accuracy under illumination-variant environments by using the LDP image and 2D-PCA algorithm. The LDP image is derived from the edge response values in different eight directions Next, the LDP image is directly inputted in 2D-PCA algorithm and nearest neighbor classifier is applied to recognize unknown user. Remark that the proposed face recognition system is very different approach when compared to previous works, because

most of previous works were used the local pattern descriptors to extract the histogram features. However, we utilize the transformed image from local pattern descriptor, i.e. LDP image as input image for further feature extraction procedure, i.e. 2D -PCA algorithm. The advantage of the proposed approach is that the illumination effects on face can be degraded by using binary pattern descriptor, and also 2D-PCA is more robust against illumination variation than global features such as PCA and LDA since 2D -PCA is a line-based local feature. In fact, we will be show that the recognition accuracy of the proposed system outperforms that of conventional approaches in the experimental results

2.1 Local Directional Pattern

The LBP operator labels the pixels of an image by thresholding a (3x3) neighborhood of each pixel with the center value and considering the results as a binary number, of which the corresponding decimal number is used for labeling. The derived binary numbers are called local binary patterns or LBP codes. While the LBP operator uses the information of intensity changes around pixels, LDP operator use the edge response values of neighborhood pixels and encode the image texture. The LDP is computed as follow. The LDP assigns an 8 bit binary code to each pixel of an input image. This pattern is then calculated by comparing the relative edge response values of a pixel by using Kirsch edge detector. Given a central pixel in the image, the eight-directional edge response values mi (i = 0, 1, ….,7) are computed by Kirsch masks as shown in Figure 1. Since the presence of a corner or an edge shows high response values in some particular directions, thus, most prominent directions of k number with high response values are selected to generate the LDP code. In other words, top-k directional bit responses, bi , are set to 1, and the remaining (8 - k) bits are set to 0. Finally, the LDP code is derived by

$$LDP_k = \sum_{i=0}^{7} b_i(m_i - m_k) \times 2^i, \quad b_i(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0, \end{cases} \qquad (1)$$

where m_k is the k^{th} most significant directional response. Figure 2 shows an example of LDP code with k-3.

m3	m2	m1
m4	X	m0
m5	m6	m7

b3	b2	b1
b4	X	b0
b5	b6	b7

Figure 1. Edge Response and LDP Binary Bit Positions

85	32	26
53	50	10
60	38	45

313	97	503
537	X	399
161	97	303

0	0	1
1	X	1
0	0	0

Figure 2. LDP Code with (k=3)

2.2 2-D Principal Component Analysis

Principal component analysis is a well-known feature extraction and data representation technique widely used in the areas of pattern recognition, computer vision, signal processing, and so on. The central underlying concept is to reduce the dimensionality of a data set while retaining the variations in the data set as much as possible. In the PCA-based face recognition method, 2D face image matrices must be previously transformed into 1D image vectors column by column or row by row fashions. However, concatenating 2D matrices into 1D vector often leads to a high-dimensional vector space, where it is difficult to evaluate the covariance matrix accurately due to its large size. Furthermore, computing the eigenvectors of a large covariance matrix is very time-consuming.

To overcome these problems, a new technique called 2D-PCA was proposed, which directly computes eigenvectors of the so-called image covariance matrix without matrix-to-vector conversion. Because the size of the image covariance matrix is equal to the width of images, which is quite small compared with the size of a covariance matrix in PCA, 2D-PCA evaluates the image covariance matrix more accurately and computes the corresponding eigenvectors more efficiently than PCA. It was reported that the recognition accuracy of 2D-PCA on several face databases was higher than that of PCA, and the feature extraction method of 2D-PCA is computationally more efficient than PCA. Unlike PCA, which treats 2D images as 1D image vectors, 2D-PCA views an image as a matrix. Consider an m by n image matrix A. Let $\in R^{(n \times d)}$ be a matrix with orthonormal columns, $n \geq d$. Projecting A onto X yields m by d matrix Y= AX. In 2D-PCA, the total scatter of the projected samples is used to determine a good projection matrix X . Suppose that there are M training face images, denoted *m* by *n* matrices Ak (k =1, 2, …,M), and the average image is denoted as $\overline{A} = 1/M \sum_k A_k$. Then, the image covariance matrix, G is given by

$$G = \frac{1}{M} \sum_{k=1}^{M} (A_k - \overline{A})^T (A_k - \overline{A}) \qquad (2)$$

It has been proven that the optimal value for the projection matrix X opt is composed by the orthonormal eigenvectors X1, X2 , …, Xd of G corresponding to the d largest eigenvalues, i.e., Xopt = [X1, X2 ,…, Xd]. Since the size of G is only n by n, computing its eigenvectors is very efficient. The optimal projection vectors of 2D-PCA, X1, X2,…,Xd are used for feature extraction. For a given face image A, the feature vector Y = [Y1, Y2, …,Yd] , in which Y has a dimension of m by d , is obtained by projecting the images into the eigenvectors as follows:

$$Y_k = (A - \overline{A}) X_k, \quad k = 1,2,\ldots d \qquad (3)$$

After feature extraction by 2D-PCA, the Euclidean distance is used to measure the similarity between the training and test features. Suppose that each training image Ak is projected onto X opt to obtain the respective 2D-PCA feature F k. Also, let A be a given image for testing and its 2D-PCA feature be F. Then, the Euclidean distance between F and F k is computed by

$$d(F, F^k) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{d}(f_{i,j}^k - f_{i,j})^2} \quad (4)$$

where k is 1,2,…M , and M is the total number of training images. This distance measurement between 2D-PCA features is further employed to classify unknown user.

3. EXPERIMENTAL RESULTS

To evaluate the robustness of the proposed method, we used images from the Yale B database and CMU-PIE database. In the Yale B database, we employ 2,414 face images for 38 subjects representing 64 illumination conditions under the frontal pose, in which subjects comprised 10 individuals in the original Yale face database B and 28 individuals in the extended Yale B database. The CMU-PIE database contains more than 40,000 facial images of 68 individuals, 21 illumination conditions, 22 light conditions, 13 poses and four different expressions. Among them, we selected each illumination and light images of 68 individuals with frontal pose (c27). So, the CMU-PIE illumination set consists of 21 images of 68 individuals (21x68 images in total), and the CMU-PIE illumination set also consists of 22 images of 68 individuals (22x68 images in total). All face images of two databases were converted as grayscale and were cropped and normalized to a resolution of (48x42) pixels. Figure 3 show an example of raw, histogram equalization, LBP, and LDP images in CMU-PIE illumination database, respectively. Remark that LDP images are divided into different groups as k number. The performance evaluation was carried out using each database of the Yale B database and CMU-PIE illumination/light database with each pre-processing images.

3.1 Yale B Database

To evaluate the performance of the proposed method, we partitioned the Yale B database into training and testing sets. Each training set comprised of seven images per subject, and the remaining images were used to test the proposed method. We selected the illumination-invariant images for training, and the remaining images with varying illumination were employed for testing. Next, we investigated the recognition performance of proposed approach with conventional recognition algorithms such as PCA and Gabor-wavelet based on LBP. For the Yale B database, the recognition results in terms of different pre-processing images and algorithms are shown in Figure 4. To further disclose the relationship between the recognition rate and dimensions of feature vectors, we showed the recognition results along with different dimensions in Figure 4. Also, we summarized the maximum recognition rates as various approaches in Table 1. As a result, the proposed approach using LDP and 2D-PCA shows a maximum recognition rate of 96.43%, when k is 3. However, the maximum recognition rates revealed 81.34% and 69.50% for PCA and Gabor-wavelets based on LBP approaches, respectively. Consequently, the recognition accuracy of proposed method was better than that of conventional methods, and it also shows performance improvement ranging from 15.09% to 29.63% in comparison to conventional methods.

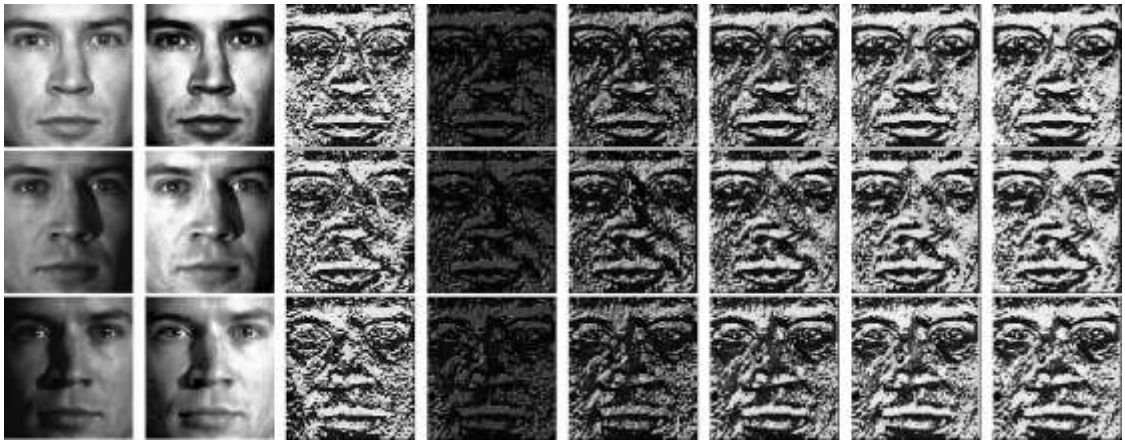


Figure 3. Input Images for CMU-PIE Illumination Database.

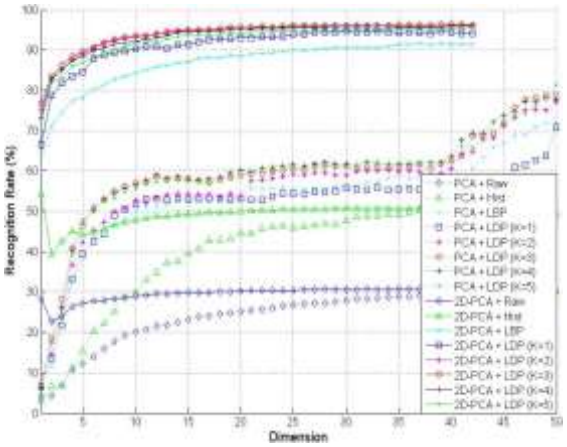


Figure 4. Recognition Rates of Yale B Database as Feature Dimensions.

Table 1. Maximum Recognition Rates on Yale B Database.

Input Images		Recognition Approaches	
	PCA	2D-PCA	Gabor-wavelets based LBP
Raw	30.03%	30.78%	57.14%
Histogram	50.61%	54.09%	69.50%
LBP	72.09%	91.54%	X
LDP (K=1)	70.77%	94.60%	X
LDP (K=2)	77.16%	95.72%	X
LDP (K=3)	78.85%	96.43%	X
LDP (K=4)	77.96%	96.10%	X
LDP (K=5)	81.34%	95.49%	X

3.2 CMU-PIE Database

For the CMU-PIE illumination/light database, each training set comprised of only one images per subject, and the remaining images were used for testing. Similar to the Yale B database, we selected an illumination-invariant image for training, and the remaining illumination-variant images were employed for testing. The recognition results for the CMU-PIE illumination database s are shown in Figure 5. For the CMU-PIE illumination database, the recognition results of various approaches shown in Table 2. In Table 2, the proposed method showed a maximum recognition rate of 100.0%, when k is 2, 3, 4, and 5, while PCA and Gabor-wavelets based on LBP approaches were 99.85% and 82.20%, respectively. As a result, the recognition accuracy of proposed method showed better performance compared to other methods, and it provide the performance improvement of 17.80% in comparison to Gabor-wavelets based on LBP approach. Similar to results of CMU-PIE illumination database, the recognition rate of proposed method showed 100.0%. Consequently, we confirmed the effectiveness of the proposed method under varying lighting conditions through these experimental results.

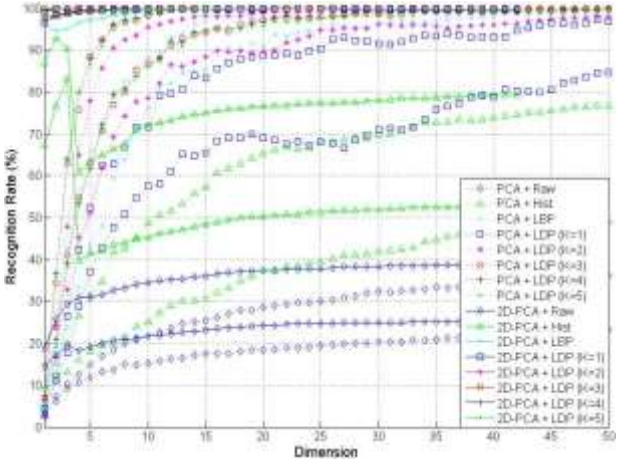


Figure 5. Recognition Rates of CMU-PIE Illumination Database as Feature Dimensions

Table 2. Maximum Recognition Rates on CMU-PIE Illumination Database.

Input Images		Recognition Approaches	
	PCA	2D-PCA	Gabor-wavelets based LBP
Raw	23.38%	25.22%	63.23 %
Histogram	49.12%	83.24%	82.20 %
LBP	98.97%	100.0%	X
LDP (K=1)	84.71%	99.93%	X
LDP (K=2)	98.09%	100.0%	X
LDP (K=3)	99.71%	100.0%	X
LDP (K=4)	99.85%	100.0%	X
LDP (K=5)	99.49%	100.0%	X

4. CONCLUSIONS

In this paper, we proposed a novel approach for achieving the illumination invariant face recognition via LDP image. Especially, we presented the face recognition methodology that utilizes the transformed image obtained from LDP as the direct input image of 2D-PCA, unlike that most of previous works used the local pattern descriptors to acquire the histogram features. The proposed method has an advantage that the illumination effects can be degraded by LDP descriptor and 2D-PCA is also more robust against illumination variation than global features. The performance evaluation was performed on the Yale B database and CMU-PIE database, and the proposed method showed the best recognition accuracy compared to different approaches. Through experimental results, we confirmed the effectiveness of the proposed method under illumination varying environments.

5. REFERENCES

[1] S. N. B. Kachare and V. S. Inamdar, Int. J. Comput. Appl., vol. 1, no. 1, 2010.

[2] T. Gong, "High-precision Immune Computation for Secure Face Recognition", International Journal of Security and Its Applications (IJSIA), vol. 6, no. 2, SERSC, pp. 293-298, 2012.

[3] L. R. Rama, G. R. Babu and L. Kishore, "Face Recognition Based on Eigen Features of Multi Scaled Face Components and Artificial Neural Network", International Journal of Security and Its Applications (IJSIA), vol. 5, no. 3, SERSC, pp. 23-44, 2012.

[4] W. Xu and E. J. Lee, "Human Face Recognition Based on Improved D-LDA and Integrated BPNNs Algorithms", International Journal of Security and Its Applications (IJSIA), vol. 6, no. 2, SERSC, pp. 121-126, 2012.

[5] M. S. Bartlett, J. R. Movellan and S. Sejnowski, "Face Recognition by Independent Component Analysis", IEEE T. Neural Networ., vol. 13, no. 6, pp. 1450-1464, 2002.

[6] S. Lawrence, C. L. Giles, A. C. Tsoi and A. D. Back, "Face recognition: A convolutional neural-network approach", IEEE T. Neural Networ., vol. 8, no. 1, pp. 98-113, 1997.

[7] W. Chen, M. J. Er and S. Wu, "Illumination compensation and nor-malization for robust face recognition using discrete cosine transform in logarithm domain", IEEE T. Syst. Man Cy. B., vol. 36, no. 2, pp. 458-466, 2006.

[8] C. Sanderson and K. K. Paliwal, "Fast features for face authentication under illumination direction changes", Pattern Recogn. Lett., vol. 24, no. 14, pp. 2409-2419, 2003.

[9] R. Basri and D. W. Jacobs, "Illumination Modeling for Face Recognition", IEEE T. Pattern Anal., vol. 25, no. 2, pp. 89-111, 2003.

[10] C. Shan, S. Gong and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study", Image Vision Comput., vol. 27, no. 6, pp. 803-816, 2009.

[11] T. Jabid, M. H. Kabir, and O. S. Chae, "Robust Facial Expression Recognition Based on Local Directional Pattern", ETRI Journal, vol. 32, no. 5. pp. 784-794, 2010.

[12] A. Georghiades, P. Belhumeur and D. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 643-660, 2001.

[13] T. Sim, S. Baker and M. Bsat, "The CMU Pose, Illumination, and Expression Database", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 12, pp. 1615-1618, 2003.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 774 - 777, 2014, ISSN:- 2319–8656

Efficient Resource Management Mechanism with Fault Tolerant Model for Computational Grids

R. Kohila
Department of Computer Science and Engineering
V.S.B Engineering College
Tamilnadu, India.

Abstract- Grid computing provides a framework and deployment environment that enables resource sharing, accessing, aggregation and management. It allows resource and coordinated use of various resources in dynamic, distributed virtual organization. The grid scheduling is responsible for resource discovery, resource selection and job assignment over a decentralized heterogeneous system. In the existing system, primary-backup approach is used for fault tolerance in a single environment. In this approach, each task has a primary copy and backup copy on two different processors. For dependent tasks, precedence constraint among tasks must be considered when scheduling backup copies and overloading backups. Then, two algorithms have been developed to schedule backups of dependent and independent tasks. The proposed work is to manage the resource failures in grid job scheduling. In this method, data source and resource are integrated from different geographical environment. Fault-tolerant scheduling with primary backup approach is used to handle job failures in grid environment. Impact of communication protocols is considered. Communication protocols such as Transmission Control Protocol (TCP), User Datagram Protocol (UDP) which are used to distribute the message of each task to grid resources.

Key Words: Grid Computing, Primary Backup, Communication Protocols, TCP- Transmission Control Protocol, UDP- User Datagram Protocol.

1. INTRODUCTION

1.1 Grid Computing

Grid Computing is distributed; large scale cluster grid computing has emerged as the next-generation parallel and distributed computing methodology, which aggregates dispersed heterogeneous resources for solving various kinds of large-scale parallel applications in science, engineering and commerce. It can integrate and utilize heterogeneous computational resources from different networks or regional areas into a high performance computational platform and can solve complex computing-intensive problems efficiently. Grid service represents convergence between high performance computation and web service. Grid aims ultimately to turn the global network of computers into a vast computational resource.

1.2 Grid Computing Overview

A distributed heterogeneous computing system consists of a distributed suite of different high-performance machines, interconnected by the high-speed networks, to perform different computationally intensive applications that have various computational requirements. Heterogeneous computing systems range from diverse elements or paradigms within a single computer to a cluster of different types of personal computers to coordinate geographically distributed machines with different architectures. Job scheduling is one the major difficult tasks in a computational grid.

2. RELATED WORK

2.1 Scheduling

Lan Foster and Car Kesselman (2004) [3] develop a fault tolerant job scheduling strategy in order to tolerate faults gracefully in an economy based grid environment. They propose a novel adaptive task check pointing based fault tolerant job scheduling strategy for an economy based grid. They present a survey with the grid community. The survey reveals that, users have to be highly involved in diagnosing failures, that most failures are due to configuration problems and that solutions for dealing with failures are mainly application-dependent.

2.2 Heuristic Algorithms

Heuristic algorithms are used for the static and dynamic tasks assignment problem. Many of these algorithms apply only to the special case where the tasks are independent i.e. with no precedence constraints. Heuristic scheduling algorithms are used in heterogeneous computing environments. These algorithms use historical data of execution time and system load and explicit constraints to schedule jobs.

2.3 Non-Evolutionary Random Scheduling Algorithm

Non-evolutionary random scheduling (RS) algorithm is used for efficient matching and scheduling of inter-dependent tasks in a distributed heterogeneous computing (DHC) system. RS is a succession of randomized task orderings and a heuristic mapping from task order to schedule. Randomized task ordering is effectively a topological sort where the outcome may be any possible task order for which the task precedent constraints are maintained.

2.4 Fault Tolerant Dynamic Scheduling Algorithm

Manimaran and Murthy (1997) [4] proposed an algorithm for dynamically scheduling arriving real-time tasks with resource and primary-backup-based fault-tolerant requirements in a multiprocessor system. This algorithm can tolerate more than one fault at a time and employs techniques such as distance concept, flexible backup overloading and resource reclaiming to improve the guarantee ratio of the system.

They address the problem of building a reliable and highly-available grid service by replicating the service on two or more hosts using the primary-backup approach. The primary goal is to evaluate the ease and efficiency with which this can be done, by first designing a primary-backup protocol using Open Grid Services Infrastructure (OSGI).

2.5 Primary-Backup Approach

Primary-backup approach, also called passive replication strategy. In this approach a backup is executed when its primary cannot complete execution due to processor failure. It does not require fault diagnosis and is guaranteed to recover all affected tasks by processor failure. Most works using the primary-backup approach consider scheduling of independent tasks.

2.5.1 Backup Overloading and Overlapping

Backup overloading is used to reduce replication cost of independent task which allows scheduling backups of multiple primaries on the same or overlapping time interval on a processor.

In Backup Overlapping, for example, two primary copies are scheduled on processor 1 and processor 3 and their backups are scheduled in an overlapping manner on processor 2.

2.6 Backup Schedules

After the earliest possible start time for a backup on all processor is determined, the time window that this backup can be scheduled on all processor is determined which is between this time and its deadline. Primary schedules and non over loadable backup schedules that are scheduled on the time window can be identified. These backup schedules could be scheduled for independent tasks or dependent tasks as interleaving technique is allowed.

3 PROPOSED WORK

The proposed system integrates resource and data source from different geographical environment. In this system, location of resource and data source is identified. There exist a fault-detection mechanism such as fail-signal and acceptance

test to detect processor and task failures. If a failure is detected in the primary, the backup will execute. Backup resources are designed with replication factors. Impact of communication protocols is considered. Communication protocols are used to distribute the message of each task to grid resource.

3.1 Scheduling Strategies

The resources and data source are managed from different environment. The location of resources and data sources is identified. There exist a fault-detection mechanism such as fail-signal and acceptance test to detect processor and task failures. If a failure is detected in the primary, the backup will execute. Backup resources are designed with replication factors. Backup overloading is used for scheduling backups of multiple primaries on the same or overlapping time interval on a processor. Resource reclaiming is also invoked when the primary completes earlier than its estimated execution time. It is necessary so that the backup slot can be released timely for new tasks.

MRC-ECT algorithm is used to schedule the backup of independent job. MCT-LRC algorithm is used to schedule the backup of dependent job. For independent tasks, scheduling of backups is independent and backups can overload as long as their primaries are schedule on different processors. Backup scheduling and overloading of dependent tasks are nontrivial and the constraint is that the backup of second task can only start after backup of first task finishes and must not be schedule on the processor where primary of the first task is located.

3.1.1 MRC-ECT Algorithm

MRC-ECT algorithm is used for scheduling backup of independent tasks. The objective is to improve resource utilization. For all processor besides the one where the primary is scheduled on, boundary schedules within the time window are considered and their replication cost is compared. This algorithm first considers the left boundary schedules of the time window. It is guaranteed to find an optimal backup schedule in terms of replication cost for a task.

3.1.2 MCT-LRC Algorithm

MCT-LRC algorithm is used for scheduling backup of independent tasks. The objective is to reduce job rejection. For all processor besides the one where the primary is scheduled on, boundary schedules within the time window are considered and the boundary schedule which can complete earliest is chosen. This algorithm first considers the left boundary schedules of the time window. Then, all existing schedules within or overlapping with the time window are examined one by one. The algorithm calculates replication cost of the earliest schedule on the current processor and records it.

3.2 Communication Protocols

Different communication protocols are used in grid environment. Transmission Control Protocol (FTP) and User Datagram Protocol (UDP) are used for data transmission. Grid File Transfer Protocol (FTP) is used for data transmission. It is used to transfer files in parallel manner. These protocols are used to distribute the message of each task to grid resource. The system analyses the data transmission in task failures.

4. CONCLUSION

In this paper, for Grid systems, we addressed the problem of fault- tolerant scheduling of jobs in heterogeneous environment. We considered the impact of communication protocols. Algorithms MRC-ECT and MCT-LRC for independent and dependent tasks respectively do not require sampling. These algorithms can schedule backups in a much faster way in heterogeneous environment. .

REFERENCES

[1] Aikebaier.A, Makoto Takizawa, Abawajy. J.H (2004), "Fault-Tolerant Scheduling Policy for Grid Computing Systems" Proceeding on Parallel and Distributed Processing Symposium (IPDPS)..

[2] Al-Omari, R., Somani, A.K., and Manimaran, G. (2001),"A New Fault-Tolerant Technique for improving Schedulability in Multiprocessor Real-Time Systems" Proceedings on Parallel Distributed Processing Symposium (IPDPS).

[3] Foster .I and Kesselman .C (2004),"Grid: Blueprint for a Future Computing Infrastructure". Morgan Kaufmann

[4] Subbiah A. and Blough D. (2004),"Distributed Diagnosis in Dynamic Fault Environments" Parallel and Distributed Systems.

[5] Qin X. and Jiang H. (2006),"A Novel Fault tolerant Scheduling Algorithm for precedence constrained Tasks in Real-Time Heterogeneous Systems" Parallel Computing.

Author:

Mrs.R.Kohila received M.E(CSE) degree from Kongu Engineering College(Affiliated to Anna University, Autonomous), Perundurai, India in 2011 and MCA degree from Bannari Amman Institute of Technology (Affiliated to Anna University), Sathyamangalam, India, in 2009 and B.Sc., degree from Trinity College of Arts and Science for women (Affiliated to Periyar University), Namakkal,India, in 2006. She has the experience in Teaching of 3+Years. Now she is currently working as an Assistant Professor in V.S.B Engineering College, Karur, Tamil Nadu, and India. His research interests include Data Mining, Advanced Data Bases, Computer Networks etc. She had presented papers in 2 National Conferences so far.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 778 - 781, 2014, ISSN:- 2319–8656

Hybrid Based Resource Provisioning in Cloud

N.Karthika
Vivekanandha College of Engineering For Women
Tiruchengode, India

K.Prabhakar
Vivekanandha College of Engineering For Women
Tiruchengode, India

R.Sangeetha
Vivekanandha College of Engineering For Women
Tiruchengode, India

Abstract: The data centres and energy consumption characteristics of the various machines are often noted with different capacities. The public cloud workloads of different priorities and performance requirements of various applications when analysed we had noted some invariant reports about cloud. The Cloud data centres become capable of sensing an opportunity to present a different program. In out proposed work, we are using a hybrid method for resource provisioning in data centres. This method is used to allocate the resources at the working conditions and also for the energy stored in the power consumptions. Proposed method is used to allocate the process behind the cloud storage.

Keywords: Cloud workload, Hybrid resource provisioning, Cloud storage and Invariant reports.

1. INTRODUCTION

Cloud Computing is the common buzzword in today's Information Technology. Cloud computing platforms are rapidly emerging as the preferred option for hosting applications in many business contexts [5]. An important feature of the cloud that differentiates it from traditional services is its apparently infinite amount of resource capacity (e.g. CPU, storage, Network) offered at a competitive rate.
It eliminates the need for setting up infrastructure which takes several months. Start-up Companies need not invest on the infrastructure because the resources are available in the cloud [6]. Cloud Computing enables users to acquire resources dynamically and elastically.

A major challenge in resource provisioning technique is to determine the right amount of resources required for the execution of work in order to minimize the financial cost from the perspective of users and to maximize the resource utilization from the perspective of service providers [4]. So, Cloud computing is one of the preferred options in today's enterprise. Resource provisioning means the selection, deployment, and run-time management of software (e.g., database management servers, load balancers) and hardware resources (e.g., CPU, storage, and network) for ensuring guaranteed performance for applications. This resource provisioning takes Service Level Agreement (SLA) into consideration for providing service to the cloud users. This is an initial agreement between the cloud users and cloud service providers which ensures Quality of Service (QoS) parameters like performance, availability, reliability, response time etc.

Based on the application needs Static Provisioning/Dynamic Provisioning and Static/Dynamic Allocation of resources have to be made in order to efficiently make use of the resources without violating SLA and meeting these QoS parameters. Over provisioning and under provisioning of resources must be avoided. Another important constraint is power consumption. Care should be taken to reduce power consumption, power dissipation and also on VM placement.

There should be techniques to avoid excess power consumption. So the ultimate goal of the cloud user is to minimize cost by renting the resources and from the cloud service provider's perspective to maximize profit by efficiently allocating the resources. In order to achieve the goal the cloud user has to request cloud service provider to make a provision for the resources either statically or dynamically so that the cloud service provider will know how many instances of the resources and what resources are required for a particular application. By provisioning the resources, the QoS parameters like availability, throughput, security, response time, reliability, performance etc must be achieved without violating SLA.

Platform as a Service is a way to rent hardware, operating systems, storage and network capacity over the internet. It delivers a computing platform or software stack as a service to run applications. This can broadly be defined as application development environment offered as a 'service' by the vendors. The development community can use these platforms to code their applications and then deploy the applications on the infrastructure provided by the cloud vendor. Here again,
the responsibility of hosting and managing the required infrastructure will be with the cloud vendor. AppEngine, Bungee Connect, LongJump, Force.com, WaveMaker are all instances of PaaS.

2. RELATED WORKS

From the last fewer, cloud computing has evolved as delivering software and hardware services over the internet. The extensive research is going on to extend the capabilities of cloud computing. Given below present related work in the area of cloud's scalability and resource provisioning in cloud computing.

In 2010 ChunyeGong, Jie Liu, Oiang Zhang, Haitao Chen and Zhenghu has discussed Characteristics of Cloud Computing. In this paper summarize the general characteristics of cloud computing which will help the development and adoption of this rapidly evolving technology. The key characteristics of cloud computing
are low cost, high reliability, high scalability, security. To make clear and essential of cloud computing, proposes the

characteristics of this area which make the cloud computing being cloud computing and distinguish it from other research area. The cloud computing has its own technical, economic, user experience characteristics. The service oriented, loose coupling, strong fault tolerant, business model and ease use are main characteristics of cloud computing. Abstraction and accessibility are two keys to achieve the service oriented conception. In loose coupling cloud computing run in a client-server model. The client or cloud users connect loosely with server or cloud providers. Strong fault tolerant stand for main technical characteristics. The ease use user experience characteristic helps cloud computing being widely accepted by non computer experts. These characteristics expose the essential of cloud computing. [1]

In 2010 Pushpendra kumar pateria, Neha Marria discussed resource provisioning in sky environment. Resource manager is used for resource provisioning and allocate of resource as user request. Offer the rule based resource manager in sky environment for utilization the private cloud resource and security requirement of resource of critical application and data .Decision is made on the basis of rule. Performance of resource manager is also evaluated by using cloudsim on basis of resource utilization and cost in sky environment. Set priorities request and allocate resource accordingly. Sky computing provides computing concurrent access to multiple clouds according user requirement. Define the Cloud services like Software as a service (SaaS), Platform as a Service (PaaS) and Infrastructure as a service. [2]

In 2010 Zhang Yu Hua, Zhang Jian ,Zhang Wei Hua present argumentation about the intelligent cloud computing system and Data warehouse that record the inside and outside data of Cloud Computing System for data analysis and data mining. Management problem of CCS are: balance between capacity and demand, capacity development planning, performance optimization, system safety management. Architecture of the Intelligence cloud computing system is defined with Data source, data warehouse and Cloud computing management information system. [3]

In 2008 discussed about the Phoenix by Jianfeng Zhan, Lei Wang, Bipo Tu, Yong Li, Peng Wang, Wei Zhou and Dan Meng. In this paper discuss the designed and implemented cloud management system software Phoenix Cloud. Different department of large organization often maintain dedicate cluster system for different computing loads. The department from big organization have operated cluster system with independent administration staffs and found many problem like resource utilization rates of cluster system are varying, dedicated cluster systems cannot provision enough resources and number of administration staff for cluster system is high. So here designed and implemented cloud management system software Phoenix Cloud to consolidate high performance computing jobs and Web service application on shared cluster system. Phoenix Cloud decreases the scale of required cluster system for a large organization.
improves the benefit of scientific computing department,
and provisions resources. [4]

In 2010 Shu-Ching Wang, Kuo-Qin Yan, Wen-Pin Liao and Shun-Sheng Wang discussed about Load Balancing in Three-Level Cloud Computing Network. Cloud computing utilize low power host to achieve high reliability. In this Cloud computing is to utilize the computing resources on the network to facilitate the execution of complicated tasks that require large-scale computation. Use the OLB scheduling algorithm is used to attempt each node keep busy and goal of load balance. Also proposed LBMM (Load Balance Min-Min) scheduling algorithm can make the minimum execution time of each task on cloud computing environment and this will improve the load unbalance of the Min-Min. In order to reach load balance and decrease execution time for each node in the three-level cloud computing network, the OLB and LBMM scheduling algorithm are integrated. The load balancing of three-level cloud computing network is utilized all calculating result could be integrated first by the secondlevel node [5]

In January 31, 2011, Sivadon Chaisiri, Bu-Sung Lee, and Dusit Niyato discuss about the Optimization of Resource Provisioning Cost. Under the resource provisioning optimal cloud provisioning algorithm illustrates virtual machine management that consider multiple provisioning stages with demand price uncertainty. In this task system model of cloud computing environment has been thoroughly explained using various techniques such as cloud consumer, virtual machine and cloud broker in details. [8]

The agent-based adaptive resource allocation is discussed in 2011 by the Gihun Jung, Kwang Mong Sim. In this paper the provider needs to allocate each consumer request to an appropriate data center among the distributed data centers because these consumers can satisfy with the service in terms of fast allocation time and execution response time. Service provider offers their resources under the infrastructure as a service model. For IaaS the service provider delivers its resources at the request of consumers in the form of VMs. To find an appropriate data center for the consumer request, propose an adaptive resource allocation model considers both the geographical distance between the location of consumer and datacenters and the workload of data center. With experiment the adaptive resource allocation model shows higher performance. An agent based test bed designed and implemented to demonstrate the proposed adaptive resource allocation model. The test bed implemented using JAVA with JADE (Java Agent Development framework). [9]

3. SYSTEM ARCHITECTURE

Dynamically adjusting the number of machines has each type to minimize total energy consumption and performance penalty in terms of scheduling delay. In my proposed using the hybrid method for resource provisioning in data centers. This method is used to allocate the resources at the working conditions and also energy stored for the power consumptions. Proposed method is used to allocate the process behind the cloud storage.

3.1 User Interface Design

In this module we design the windows for the project. These windows are used to send a message from one to another. In this module mainly we are focusing the login design page with the Partial knowledge information. Application Users need to view the application they need to login through the User Interface GUI is the media to connect User and Media Database.

3.2 Dynamic Capacity Provisioning In Data Centers

In this section we address the simulation of heterogeneous active machine. Here we are going to create difference machine for stored based on the client demand that is production data centers often comprise heterogeneous machines with different capacities and energy consumption characteristics..The energy level consumption is updated by the cloud service provider which belongs to the datacenter. Data center have more right the route the heterogeneous active machine. This area regulates Heterogeneous active machine creation. Also it regulates memory consumption and key challenge that often has been overlooked or considered difficult to address is heterogeneity,

3.3 Machine Heterogeneity Model Approach

This simulation addresses the production data centers often comprise several types of machines from multiple update. They have heterogeneous processor architectures and speeds, hardware features, memory and disk capacities. Consequently, they have different runtime energy consumption rates. Here we are going to create difference machine for stored based on the client demand that is production data centers often comprise heterogeneous machines with different capacities and energy consumption characteristics..The energy level consumption is updated by the cloud service provider which belongs to the datacenter.

3.4 Resource Monitoring and Management System

Production data centers receive a vast number of heterogeneous resource requests with diverse resource demands, durations, priorities and performance. The heterogeneous nature of both machine and workload in production cloud environments has profound implications on the design of DCP schemes. Here we address accurate characterization of both workload and machine heterogeneities. Standard K-means clustering, we show that the heterogeneous workload can be divided into multiple task classes with similar characteristics in terms of resource and performance objectives.

3.5 Dynamic Capacity Provisioning Approach

The workload traces contain scheduling events, resource demand and usage records. The job is an application that consists of one or more tasks. Each task is scheduled on a single physical machine. When a job is submitted, the user can specify the maximum allowed resource demand for each task in terms of required CPU and memory size. Dynamically adjust the number of active machines in a data center in order to reduce energy consumption while meeting the service level objectives (SLOs) of workloads. The coordinates of each point in these figures correspond to a combination of CPU and memory requirements.

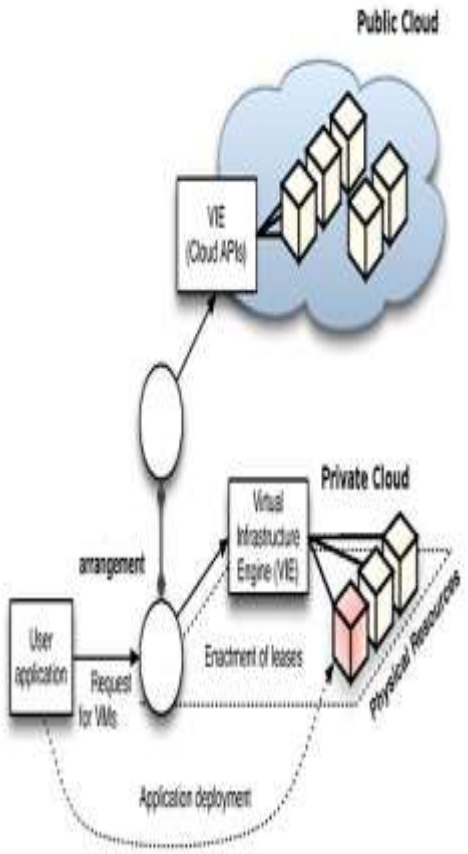


Figure 1. Architecture of proposed framework

The resource provisioning in Cloud Computing, the long-held dream of computing as a utility, has the potential to transform a large part of the IT industry, making software even more attractive as a service and shaping the way IT hardware is designed and purchased. Developers with innovative ideas for new Internet services no longer require the large capital outlays in hardware to deploy their service or the human expense to operate it. They need not be concerned about over-provisioning for a service whose popularity does not meet their predictions, thus wasting costly resources, or under-provisioning for one that becomes wildly popular, thus missing potential customers and revenue. The Methodology based on Infrastructure as a Service layer to access resources on-demand. A Rule Based Resource Manager is proposed to scale up private cloud and presents a cost effective solution in terms of money spent to scale up private cloud on-demand by taking public cloud's resources and that never permits secure information to cross the organization's firewall in hybrid cloud. Also set the time for public cloud and private cloud to fulfill the request.

4. CONCLUSION

The user's usages have large number of progress in an environment. So there have large number of problems are occurred in the cloud. The resource provisioning can be overcome by hybrid method. This proposed method is used to allocate the resources with working conditions. It shows the energy is very efficiency and the overcome the workload with the good performance.

6. REFERENCES

[1]http://www.youtube.com/yt/press/statistics.html
[2]http://nlp.stanford.edu/software/corenlp.shtml
[3]Collins English Dictionary, entry for "lemmatise"
[4]L. Ratinov and D. Roth, Design Challenges and Misconceptions in Named Entity Recognition. CoNLL (2009)
[5]G. A. Miller.Wordnet: A lexical database for english. (11):39-41.
[6]Chengde Zhang, Xiao Wu, Mei-Ling Shyu and QiangPeng, " Adaptive Association Rule Mining for Web Video Event Classification", 2013 IEEE 14th International Conference on Information Reuse and Integration (IRI), page 618-625.
[7] Y. Song, M. Zhao, J. Yagnik, and X. Wu.Taxonomic classification for web-based videos.In CVPR, 2010.
[8] Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li. Youtube-cat: Learning to categorize wild web videos. In CVPR, 2010.
[9] http://www.ranks.nl/resources/stopwords.html
[10]http://cs.nyu.edu/grishman/jet/guide/PennPOS.html
[11]Roth and D. Zelenko, Part of Speech Tagging Using a Network of Linear Separators. Coling-Acl, The 17th International Conference on Computational Linguistics (1998) pp. 1136—1142
[12]O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Au- tomatic annotation of human actions in video. In Proc. of ICCV, 2009.
[13]Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld.Learning realistic human actions from movies. In Proc. of CVPR, 2008
[14]M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy automatic naming of characters in tv video. In Proc. of BMVC, 2006.
[15]F. Smeaton, P. Over, and W. Kraaij.Evaluation campaigns and trecvid. In Proc. of ACM Workshop on Multimedia In- formation Retrieval, 2006
[16]J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In Proc. of ACM MM, 2007.
[17] M. E. Sargin, H. Aradhye, P. J. Moreno, and M. Zhao. Au- diovisual celebrity recognition in unconstrained web videos. In Proc. of ICASSP, 2009.
[18] J. Liu, J. Luo, and M. Shah.Recognizing realistic actions from videos.In Proc. of CVPR, 2009.
[19] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569-571, Nov. 1999

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 782 - 784, 2014, ISSN:- 2319–8656

Authentic Data Access Scheme for Variant Disruption-Tolerant Networks

S.Raja Rajeshwari
Vivekanandha College Of Engineering For Women
Tiruchengode, India

K. Prabhakar
Vivekanandha College Of Engineering For Women
Tiruchengode, India

S.Fowjiya
Vivekanandha College Of Engineering For Women
Tiruchengode, India

Abstract: Mobile nodes in military environments such as a battlefield or a hostile region are likely to suffer from intermittent network connectivity and frequent partitions. Disruption-tolerant network (DTN) technologies are becoming successful solutions that allow wireless devices carried by soldiers to communicate with each other and access the confidential information or command reliably by exploiting external storage nodes. However, the problem of applying CP-ABE in decentralized DTNs introduces several security and privacy challenges with regard to the attribute revocation, key escrow, and coordination of attributes issued from different authorities. In this paper, we propose a secure data retrieval scheme using CP-ABE for decentralized DTNs where multiple key authorities manage their attributes independently. We demonstrate how to apply the proposed mechanism to securely and efficiently manage the confidential data distributed in the disruption-tolerant military network. Since some users may change their associated attributes at some point (for example, moving their region), or some private keys might be compromised, key revocation (or update) for each attribute is necessary in order to make systems secure. This implies that revocation of any attribute or any single user in an attribute group would affect the other users in the group. It may result in bottleneck during rekeying procedure, or security degradation due to the windows of vulnerability if the previous attribute key is not updated immediately**.**

Keywords: component; formatting; style; styling; insert (Minimum 5 to 8 key words)

1. INTRODUCTION

We ask that authors follow some simple guidelines. This document is a template. An electronic copy can be downloaded from the journal website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website

Delay-tolerant networking (DTN) is an approach to computer network architecture that seeks to address the technical issues in heterogeneous networks that may lack continuous network connectivity. Examples of such networks are those operating in mobile or extreme terrestrial environments, or planned networks in space.

Recently, the term disruption-tolerant networking has gained currency in the United States due to support from DARPA, which has funded many DTN projects. Disruption may occur because of the limits of wireless radio range, sparsity of mobile nodes, energy resources, attack, and noise.

Roy [4] and Chuah [5] introduced storage nodes in DTNs where data is stored or replicated such that only authorized mobile nodes can access the necessary information quickly and section of confidential data including access control methods that are cryptographically enforced [6], [7]. In many cases, it
is desirable to provide differentiated access services such that data access policies are defined over user attributes or roles, which are managed by the key authorities.
In this case, it is a reasonable assumption that multiple key authorities are likely to manage their own dynamic attributes for soldiers in their deployed regions or echelons, which could be frequently changed (e.g., the attribute representing current location of moving soldiers) [4], [8], [9]. We refer to this DTN architecture where multiple authorities issue and manage their own attribute keys independently as a decentralized DTN [10].

The concept of attribute-based encryption (ABE) is a promising approach that fulfills the requirements for secure data retrieval in DTNs. ABE features a mechanism that enables an access control over encrypted data using access policies and ascribed attributes among private keys and ciphertexts. Especially, ciphertext-policy ABE (CP-ABE) provides a scalable way of encrypting data such that the encryptor defines the attribute set that the decryptor needs to possess in order to decrypt the ciphertext [13]. Thus, different users are allowed to decrypt different pieces of data per the security policy.

2. RELATED WORKS

In CP-ABE, the ciphertext is encrypted with an access policy chosen by an encryptor, but a key is simply created with respect to an attributes set. CP-ABE is more appropriate to DTNs than KP-ABE because it enables encryptors such as a commander to choose an access policy on attributes and to encrypt confidential data under the access structure via encrypting with the corresponding public keys or attributes [4], [7], [15].

Most of the existing ABE schemes are constructed on the architecture where a single trusted authority has the power to generate the whole private keys of users with its master secret information [11], [13], [14]. Thus, the key escrow problem is inherent such that the key authority can decrypt every ciphertext addressed to users in the system by generating their secret keys at any time. Chase *et al.* presented a distributed KP-ABE scheme that solves the key escrow problem in a multiauthority system. In this approach, all (disjoint) attribute authorities are participating in the key generation protocol in a distributed way such that they cannot pool their data and link multiple attribute sets belonging to the same user. One

disadvantage of this fully distributed approach is the performance degradation. Since there is no centralized authority with master secret information, all attribute authorities should communicate with each other in the system to generate a user's secret key.

3. SYSTEM DESIGN

3.1 Existing System

When multiple authorities manage and issue attribute keys to users independently with their own master secrets, it is very hard to define fine-grained access policies over attributes issued from different authorities.

The problem of applying the ABE to DTNs introduces several security and privacy challenges. Since some users may change their associated attributes at some point (for example, moving their region), or some private keys might be compromised, key revocation (or update) for each attribute is necessary in order to make systems secure. However, this issue is even more difficult, especially in ABE systems, since each attribute is conceivably shared by multiple users (henceforth, we refer to such a collection of users as an attribute group)

Another challenge is the key escrow problem. In CP-ABE, the key authority generates private keys of users by applying the authority's master secret keys to users' associated set of attributes. The last challenge is the coordination of attributes issued from different authorities. When multiple authorities manage and issue attributes keys to users independently with their own master secrets, it is very hard to define fine-grained access policies over attributes issued from different authorities.

3.2 Proposed System

First, immediate attribute revocation enhances backward/forward secrecy of confidential data by reducing the windows of vulnerability.

Second, encryptors can define a fine-grained access policy using any monotone access structure under attributes issued from any chosen set of authorities.

Third, the key escrow problem is resolved by an escrow-free key issuing protocol that exploits the characteristic of the decentralized DTN architecture. The key issuing protocol generates and issues user secret keys by performing a secure two-party computation (2PC) protocol among the key authorities with their own master secrets. The 2PC protocol deters the key authorities from obtaining any master secret information of each other such that none of them could generate the whole set of user keys alone.

Thus, users are not required to fully trust the authorities in order to protect their data to be shared. The data confidentiality and privacy can be cryptographically enforced against any curious key authorities or data storage nodes in the proposed scheme.

3.2.1 Data confidentiality:

Unauthorized users who do not have enough credentials satisfying the access policy should be deterred from accessing the plain data in the storage node. In addition, unauthorized access from the storage node or key authorities should be also prevented.

3.2.2 Collusion-resistance:

If multiple users collude, they may be able to decrypt a ciphertext by combining their attributes even if each of the users cannot decrypt the ciphertext alone.

3.2.3 Backward and forward Secrecy

In the context of ABE, backward secrecy means that any user who comes to hold an attribute (that satisfies the access policy) should be prevented from accessing the plaintext of the previous data exchanged before he holds the attribute. On the other hand, forward secrecy means that any user who drops an attribute should be prevented from accessing the plaintext of the subsequent data exchanged after he drops the attribute, unless the other valid attributes that he is holding satisfy the access policy.

Please use a 9-point Times Roman font, or other Roman font with serifs, as close as possible in appearance to Times Roman in which these guidelines have been set. The goal is to have a 9-point text, as you see here. Please use sans-serif or non-proportional fonts only for special purposes, such as

.

4. SYSTEM IMPLEMENTATION

4.1 Key Authorities

They are key generation centers that generate public/secret parameters for CP-ABE. The key authorities consist of a central authority and multiple local authorities. We assume that there are secure and reliable communication channels between a central authority and each local authority during the initial key setup and generation phase. Each local authority manages different attributes and issues corresponding attribute keys to users.

They grant differential access rights to individual users based on the users' attributes. The key authorities are assumed to be honest-but-curious. That is, they will honestly execute the assigned tasks in the system; however they would like to learn information of encrypted contents as much as possible.

4.2 Storage node:

This is an entity that stores data from senders and provide corresponding access to users. It may be mobile or static. Similar to the previous schemes, we also assume the storage node to be semi-trusted that is honest-but-curious.

4.3 Sender:

This is an entity who owns confidential messages or data (e.g., a commander) and wishes to store them into the external data storage node for ease of sharing or for reliable delivery to users in the extreme networking environments. A sender is responsible for defining (attribute based) access policy and enforcing it on its own data by encrypting the data under the policy before storing it to the storage node.

4.4 User

This is a mobile node who wants to access the data stored at the storage node (e.g., a soldier). If a user possesses a set of attributes satisfying the access policy of the encrypted data defined by the sender, and is not revoked in any of the attributes, then he will be able to decrypt the ciphertext and obtain the data.

5. CONCLUSION

The concept of attribute-based encryption (ABE) is a promising approach that fulfills the requirements for secure data retrieval in DTNs. ABE features a mechanism that enables an access control over encrypted data using access policies and ascribed attributes among private keys and ciphertexts. Especially, Ciphertext policy ABE (CP-ABE) provides a scalable way of encrypting data such that the encryptor defines the attribute set that the decryptor needs to possess in order to decrypt the ciphertext. Thus, different users are allowed to decrypt different pieces of data per the security policy. When multiple authorities manage and issue attribute keys to users independently with their own master secrets, it is very hard to define fine-grained access policies over attributes issued from different authorities.

5. REFERENCES

[1] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine, "Maxprop: Routing for vehicle-based disruption tolerant networks," in Proc.IEEE INFOCOM, 2006, pp. 1–11.

[2] M. Chuah and P. Yang, "Node density-based adaptive routing scheme for disruption tolerant networks," in Proc. IEEE MILCOM, 2006, pp.1–6.

[3] M. M. B. Tariq, M. Ammar, and E. Zequra, "Mesage ferry route design for sparse ad hoc networks with mobile nodes," in Proc. ACM MobiHoc, 2006, pp. 37–48.

[4] S. Roy andM. Chuah, "Secure data retrieval based on ciphertext policy attribute-based encryption (CP-ABE) system for the DTNs," Lehigh CSE Tech. Rep., 2009.

[5] M. Chuah and P. Yang, "Performance evaluation of content-based information retrieval schemes for DTNs," in Proc. IEEE MILCOM, 2007, pp. 1–7.

[6] M. Kallahalla, E. Riedel, R. Swaminathan, Q. Wang, and K. Fu, "Plutus: Scalable secure file sharing on untrusted storage," in Proc. Conf. File Storage Technol., 2003, pp. 29–42.

[7] L. Ibraimi, M. Petkovic, S. Nikova, P. Hartel, and W. Jonker, "Mediated ciphertext-policy attribute-based encryption and its application," in Proc. WISA, 2009, LNCS 5932, pp. 309–323.

[8] N. Chen, M. Gerla, D. Huang, and X. Hong, "Secure, selective group broadcast in vehicular networks using dynamic attribute based encryption," in Proc. Ad Hoc Netw. Workshop, 2010, pp. 1–8.

[9] D. Huang and M. Verma, "ASPE: Attribute-based secure policy enforcement in vehicular ad hoc networks," Ad Hoc Netw., vol. 7, no. 8, pp. 1526–1535, 2009.

[10] A. Lewko and B. Waters, "Decentralizing attribute-based encryption," Cryptology ePrint Archive: Rep. 2010/351, 2010.

[11] A. Sahai and B. Waters, "Fuzzy identity-based encryption," in Proc. Eurocrypt, 2005, pp. 457–473.

[12] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in Proc. ACM Conf. Comput. Commun. Security, 2006, pp. 89–98.

[13] J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-policy attributebased encryption," in Proc. IEEE Symp. Security Privacy, 2007, pp. 321–334.

[14] R. Ostrovsky, A. Sahai, and B. Waters, "Attribute-based encryption with non-monotonic access structures," in Proc. ACM Conf. Computer Commun. Security, 2007, pp. 195–203.

[15] S. Yu, C. Wang, K. Ren, and W. Lou, "Attribute based data sharing with attribute revocation," in Proc. ASIACCS, 2010, pp. 261–270.

[16] [16] A. Boldyreva, V. Goyal, and V. Kumar, "Identity-based encryption with efficient revocation," in Proc. ACM Conf. Comput. Commun. Security, 2008, pp. 417–426.

[17] M. Pirretti, P. Traynor, P. McDaniel, and B. Waters, "Secure attributebased systems," in Proc. ACMConf. Comput. Commun. Security, 2006, pp. 99–112.

[18] Junbeom Hur and Kyungtae Kang "Secure Data Retrieval for Decentralized Disruption-Tolerant Military Network",, IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 22, NO. 1, FEBRUARY 2014

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 785 - 790, 2014, ISSN:- 2319–8656

Reverse Engineering for Documenting Software Architectures, a Literature Review

Hind Alamin Mohamed
College of Computer Science and Information Technology,
Sudan University of Science and Technology,
SUST University, SUDAN

Hany H Ammar
Lane Computer Science and Electrical Engineering Department,
College of Engineering and Mineral Resources,
West Virginia University, USA

Abstract: Recently, much research in software engineering focused on reverse engineering of software systems which has become one of the major engineering trends for software evolution. The objective of this survey paper is to provide a literature review on the existing reverse engineering methodologies and approaches for documenting the architecture of software systems. The survey process was based on selecting the most common approaches that form the current state of the art in documenting software architectures. We discuss the limitations of these approaches and highlight the main directions for future research and describe specific open issues for research.

Keywords: Reveres Engineering; Software Architecture; Documenting Software Architectures; Architectural Design Decisions.

1. INTRODUCTION

Reverse engineering has become one of the major engineering trends for software evolution. Reverse engineering is defines as the process of analyzing an existing system to determine its current components and the relationship between them. This process extracts and creates the design information and new forms of system representations at a higher level of abstraction [1, 2]. Garg et al. categorized engineering into forward engineering and reverse engineering. Both of these types are essential in the software development life cycle. The forward engineering refers to the traditional process for developing software which includes: gathering requirements, designing and coding process till reach the testing phase to ensure that the developed software satisfied the required needs [1]. While reverse engineering defined as the way of analyzing an existing system to identify its current components and the dependencies between these components to recover the design information, and it creates other forms of system representations [1, 2].

Legacy systems are old existing systems which are important for business process. Companies rely on these legacy systems and keep them in operations [2]. Therefore, reverse engineering is used to support the software engineers in the process of analyzing and recapturing the design information of complex and legacy systems during the maintenance phase [2, 3].

In addition, the main objectives of reverse engineering are focused on generating alternative views of system's architecture, recover the design information, re-documentation, detect limitations, represent the system at higher abstractions and facilitate reuse [1, 2, 4].

The main purpose of this survey paper is to achieve the following objectives: provide a literature review on the existing reverse engineering methodologies for documenting the architecture of software systems, and highlights the open issues and the directions for future research.

The rest of the paper is organized as follows: *Section 2*; presents a literature review of the common existing researches on reverse engineering from different perspectives. *Section 3*; highlights the new research areas as open issues for future works. Finally, concludes with summarizing the main contribution and the future research.

2. LITERATURE REVIEW

Program understanding plays a vital role in most of software engineering tasks. In fact; the developers use the software documentation to understand the structure and behavior of existing systems [4, 5]. However, the main problem that developers face is that the design document or others software artifacts were out-of-date to reflect the system's changes. As a result, more effort and time needed for understanding the software rather that modifying it [4, 5]. The following sections will introduce the most common reverse engineering approaches that focused in documenting the architecture of software from different perspectives.

2.1 Reverse Engineering for Understanding Software Artifacts

Kumar explained that developers should understand the source code based on the static information and dynamic information [5]. The static information explained the structural characteristic of the system. While dynamic information explained the dynamic characteristics or behaviors of the system. Hence, these details help the developers on understanding the source code in order to maintain or evaluate the system. However, Kumar clarified that few reverse engineering tools supported both of dynamic and static information [5]. Therefore, he presented alternative methodology to extract the static and dynamic information from existing source code. This methodology focused on using one of the reverse engineering tools; namely, *Enterprise Architect (EA)* to extract the static and dynamic views.

Additionally, all of the extracted information was represented in form of Unified Modeling Language (UML) models. The main purpose was to get the complementary views of software in the form of state diagrams and communication diagrams. The stages of this methodology are summarized as it shown in Figure 1.

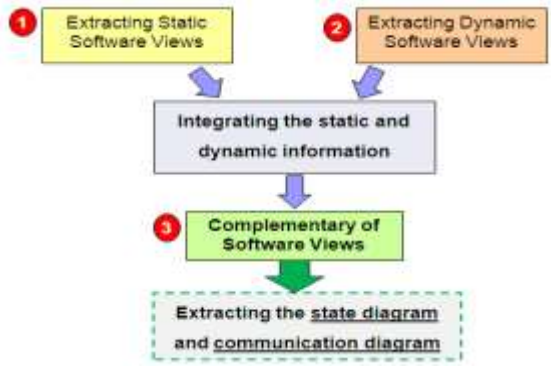


Figure 1. Reverse Engineering thorough Complementary Software Views [5]

This proposed methodology was very useful for supporting developers to understand the software artifacts of existing software systems. However, the methodology needs to support additional stakeholder beside the developers in order to identify the stakeholders' concerns and their decisions about the whole system.

2.2 Model Driven Reverse Engineering

Model driven reverse engineering (MDRE) was proposed as described in [6] to improve the traditional reverse engineering activities and legacy technologies. It is used to describe the representation of derived models from legacy systems to understand their contents. However, most of MDRE solutions focused on addressing several types of legacy system scenarios, but these solutions are not complete and they do not cover the full range of legacy systems. The work also introduced several reverse engineering processes such as: the technical/functional migration, processes of MDRE [6].

Recently, Hugo et al. presented a generic and extensible MDRE framework called "MoDisco". This framework is applicable to different refactoring and re-documentation techniques [6]. The architecture of MoDisco is represented in three layers, each layer is comprised of one or more components (see Figure 2). The components of each layers provided high adaptability because they are based on the nature of legacy system technologies and the scenario based on reverse engineering.

However, the MoDisco framework was limited to traditional technologies such as: JAVA, JEE (including JSP) and XML. This framework needs to be extended to support additional technologies and to add more advanced components to improve the system comprehension, and expose the key architecture design decisions.

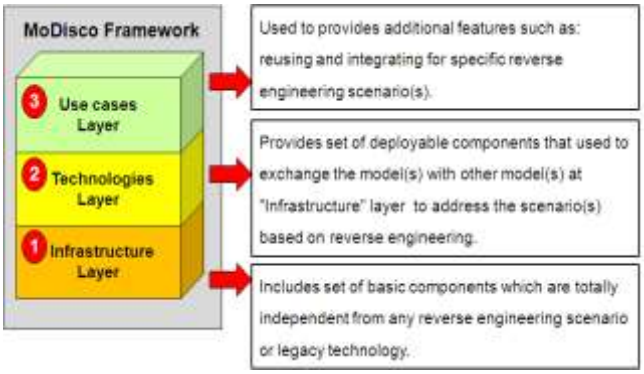


Figure 2. MoDisco Framework's Architecture [6, p9]

2.3 Documenting of Architectural Design Decisions (ADDs)

Historically, Shaw and Garlan introduced the concepts of software architecture and defined the system in terms of computational components and interactions between these components as indicated in [7]. Moreover, Perry and Wolf defined software architecture in terms of elements, their properties, and the relationships among these elements. They suggested that the software architecture description is the consequence of early design decisions [7].

Software architecture is defined by the recommended practice (ANSI/IEEE Std 1471-2000) as: the fundamental organization of a system, embodied in its components, their relationships to each other and the environment, and the principles governing its design and evolution. Software architecture development is based on a set of architectural design decisions (ADDs). This is considered as one of the important factors in achieving the functional and non-functional requirements of the system [8].

Che explained that the process of capturing and representing ADDs is very useful for organizing the architecture knowledge and reducing the possibility of missing this knowledge [8]. Furthermore, the previous research focused on developing tools and approaches for capturing, representing and sharing of the ADDs.

However, Che clarified that most of the previous research proposed different methods for documenting ADDs, and these methods rarely support architecture evaluation and knowledge evaluation in practice [8]. Accordingly, Che et al. presented an alternative approach for documenting and evaluating ADDs. This approach proposed solutions described in the following subsections [8, 9]:

2.3.1 Collecting of Architectural Design Decisions

The first solution focused on creating a general architectural framework for documenting ADDs called the Triple View Model (TVM). The framework includes three different views for describing the notation of ADDs as shown in Figure 3. It also covers the features of the architecture development process [8, 9].

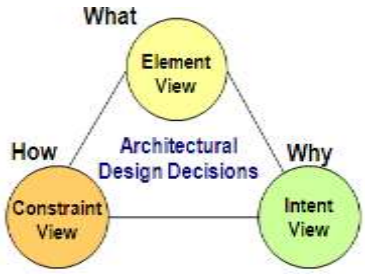


Figure 3. Triple View Model Framework [8, p1374]

As it shown in Figure 3; the *Element View* describes the elements that should be defined to develop the architecture; such as: Computation elements, Data elements, and Connector elements. The *Constraint View* explains how the elements interact with each other by defining what the system should do and not to do, the constraint(s) on each element of the element view. Additionally, define the constraints on the interaction and configuration among the elements.

Finally, the *Intent View* includes the rationale decision that made after analyzing all the available decisions, Moreover, the selection of styles and patterns for the architecture and the design of the system.

2.3.2 Scenario-Based Documentation and Evaluation Method

The second solution called SceMethod is based on the TVM framework. The main purpose is to apply the TVM framework by specifying its views through the end-user scenarios; then manage the documentation and the evaluation needs for ADDs [8, 10].

2.3.3 UML Metamodel

The third solution is focused on developing the UML Metamodel for the TVM framework. The main purpose was to make each view of TVM specified by classes and a set of attributes for describing ADD information. Accordingly, this solution provided the following features [8]: a) establish traceable evaluation of ADDs, b) apply the evaluation related to the specified attributes, c) support multiple ways on documenting during the architecture process and allow explicit evaluation knowledge of ADDs.

Furthermore, TVM and SceMethod solution was validated in using a case study to ensure the applicability and the effectiveness. Supporting the ADD documentation and evaluation in geographically separated software development (GSD) is currently work in progress.

2.4 Comparison of Existing Architectural Design Decisions Models

Researchers made a great of effort to present related tools and models for capturing, managing, and sharing the ADDs. These proposed models were based on the concept of architectural knowledge to promote the interaction between the stakeholders and improve the architecture of the system [8, 11].

Accordingly in [11], Shahin et al. presented a comparison study that is based on surveying and comparing the existing architectural design decisions models. Their comparison included nine ADD models and used six criteria based on desired features [11, 12]. The main reason was to investigate the ADD models to decide if there are similarities and differences in capturing the ADDs. Moreover, the study aimed at finding the desired features that were missed according to the architecture needs [11]. The authors in [11] classified the ADD elements into two categories: *major elements* and *minor elements*. The *major elements* refer to the consensus on capturing and documenting ADDs based on the constraints, rationale, and alternative decisions. While the *minor elements* refer to the elements that used without consensus on capturing and documenting the ADDs, such as: stakeholders, problem, group, status, dependency, artifacts, and phase/iteration.

The main observations of this comparison study are highlighted as follow: 1) all of the selected ADD models included the major elements and used different terms to express similar concepts of the architecture design. 2) Most ADD models used different minor elements for capturing and documenting ADDs. 3) All the selected ADD models deal with the architecture design as a decision making process. 4) While not all of them are supported by tools, some were based on only textual templates for capturing and documenting ADDs. 5) The most important observation was that most of existing ADD tools do not provide support for ADD personalization which refers to the ability of stakeholders to communicate with the stored knowledge of ADD [11, 12] based on their own profile.

We summarize the approaches and methodologies described in this section in Table 1. The main observation is that existing methods are focused on the developer's concerns and viewpoints as the main stakeholder. Recent approaches such as: Triple View Model (TVM) [8], scenario-based method (SceMethod) [9], and managing ADDs [10] suggested the need for alternative solutions for supporting ADDs personalization for different stakeholders.

3. OPEN ISSUES

We describe in this section the open issues that require further research based on the research work described in the previous section. These issues are listed as follows:

- There is a significant need to develop alternative approaches of reverse engineering for documenting the architectures that should simplify and classify all of the available information based on identifying the stakeholders' concerns and their decisions about the system.
- Improve the system's comprehension by establishing more advanced approaches for understanding the software artifacts. These approaches should help in documenting the architecture at different levels of abstractions and granularities based on the stakeholders concerns.
- Finally, it's important to support multiple methods and guidelines on how to use the general ADDs framework in the architecting process. These methods should be base on the architecture needs, context and challenges in order to evaluate the ADDs in the architecture development and evolution processes.

Table 1. Examples of some Methodologies and Approaches for Documenting Software Architecture

#	Author (year)	Problem Statement	Proposed Solution(s)	Results and Findings	Limitation(s)
1	**Kumar (2013)**	Reverse engineering for understanding the software artifacts	- Alternative methodology to extract the static and dynamic information from the source code. - The main purpose is to get complementary views of software systems.	- This methodology support *developers* to achieve the reverse engineering goals in order to understand the artifacts of software systems.	This methodology needs to support additional stakeholder beside the developers in order to identify the stakeholders' concerns and their decisions about the whole system.
2	**Hugo et al (2014)**	Understanding the contents of the legacy systems using model driven reverse engineering (MDRE)	- Generic and extensible MDRE framework called "MoDisco". - This framework is applicable to different types of legacy systems.	- MoDisco provided high adaptability because it is based on the nature of legacy system technologies and the scenario(s) based on reverse engineering.	MoDisco should extend to support additional technologies and include more advanced components to improve the system comprehension.
3	**Che et al (2011)**	Collecting architectural design decisions (ADDs)	- Triple View Model (TVM) an architecture framework for documenting ADDs.	- TVM framework includes three different views for describing the notation of ADDs. - TVM covers the main features of the architecture process.	TVM framework should extend to manage the evaluation and documentation of ADDs by specifying its views through the stakeholders' scenarios.
4	**Che et al (2012)**	Managing the documentation and evolution of the architectural design decisions	- Scenario based method (*SceMethod)* for documenting and evaluating ADDs. - This solution is based on TVM. The main purpose is to apply TVM for specifying its views through end-user scenario(s).	- Manage the documentation and the evaluation needs for ADDs through stakeholders' scenario(s).	There is a need to support multiple ways on managing and documenting the ADDs during the architecture process.

#	Author (year)	Problem Statement	Proposed Solution(s)	Results and Findings	Limitation(s)
5	**Che (2013)**	Documenting and evolving the architectural design decisions	- Developed UML Metamodel for the TVM framework. The main purpose was to make each view of TVM specified by classes and a set of attributes for describing ADDs information.	- Apply the evaluation related to the specified attributes and establish traceable evaluation of ADDs, - Allow explicit evaluation knowledge of ADDs. - Support multiple ways for documenting ADDs during the architecture process.	This solution is focused on the developers view point and their work is currently in progress to support the ADD documentation and evaluation in geographically separated software development (GSD).
6	**Shahin et al (2009)**	A survey of architectural design decision models and tools	- The purpose of this survey was to investigate ADD models to decide if there are any similar concepts or differences on capturing ADD. - The survey classified ADD concept into two categories: *Major elements* which refer to the consensus on capturing and documenting ADD based on the constraint, rationale and alternative of decision. While the *Minor elements* refers to the elements that used without consensus on capturing and documenting ADD. - Moreover, to clarify the desired features that are missed according to the architecture needs	- All of selected ADD models include the *major elements*. - Most of ADD models are based on using different *minor elements* for capturing and documenting the ADD. - All of selected ADD models deal with the architecture design as the decision making process. - Not all models were supported by tools. Hence, some of these ADD based on *text template* for capturing and documenting ADDs. - However, most of existing ADD tools do not support the ability of stakeholders to communicate with the stored knowledge of ADD.	There is a need to focus on stakeholder to communicate with the stored knowledge of ADDs. This could be achieved by applying the scenario based documentation and evaluation methods through stakeholders' scenario(s) to manage the documentation and the evaluation needs for ADDs.

4. CONCLUSIONS

This paper presented a survey on the current state of the art in documenting the architectures of existing software systems using reverse engineering techniques. We compared existing methods based on their findings and limitations. The main observation is that existing methods are focused on the developer's concerns and viewpoints as the main stakeholder.

We outlined several open issues for further research to develop alternative approaches of reverse engineering for documenting the architectures for development and evolution. These issues show the need to simplify and classify available information based on identifying the stakeholders' concerns and viewpoints about the system, improve comprehension by documenting the architecture at different levels of abstractions and granularities based on the stakeholders concerns, and support multiple methods and guidelines on how to use the ADDs framework based on the architecture needs, context and challenges in order to evaluate these ADDs during the architecture development and evolution processes.

5. ACKNOWLEDGMENTS

This research work was funded in part by Qatar National Research Fund (QNRF) under the National Priorities Research Program (NPRP) Grant No.: 7 - 662 - 2 - 247

6. REFERENCES

[1] Mamta Gar and Manoj Kumar Jindal. 2009. Reverse Engineering – Roadmap to Effective software Design. In Proceedings of 2th International Journal of Recent Trends in Engineering. Information Paper, vol.1, (May 2009).

[2] Rosenberg, Linda H. and Lawrence E. Hyatt, Software re-engineering. Software Assurance Technology Center, 1996. http://www.scribd.com/doc/168304435/ Software-Re-Engineering1, visited on 26 April 2014.

[3] M. Harman, W. B. Langdon and W. Weimer.2013. Genetic Programming for Reverse Engineering, In R. Oliveto and R. Robbes, editors, In Proceedings of 20th Working Conference on Reverse Engineering (WCRE'13). Koblenz, Germany (14-17 October 2013), IEEE, 2013.

[4] M. Harman, Yue Jia, W. B. Langdon, Justyna Petke, Iman H. Moghadam, Shin Yoo and Fan Wu. 2014. Genetic Improvement for Adaptive Software Engineering. In Proceedings of 9th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS'14). Hyderabad, India (2-3 June 2014), ACM, 2014.

[5] Niranjan Kumar. 2013. An Approach for Reverse Engineering thorough Complementary Software Views. In Proceedings of International Conference on Emerging Research in Computing, Information, Communication and Applications (ERCICA'13), 2013, 229-234.

[6] Hugo Brunelière, Jordi Cabot, Grégoire Dupé and Frédéric Madiot. 2014. MoDisco: A Model Driven Reverse Engineering Framework. Information and Software Technology 56, no. 8, 2014, 1012-1032.

[7] May Nicholas. 2005. A survey of software architecture viewpoint models. In Proceedings of 6th Australasian Workshop on Software and System Architectures, 2005, 13-24.

[8] Meiru Che. 2013. An Approach to Documenting and Evolving Architectural Design Decisions. In Proceedings of International Conference on Software Engineering (ICSE'13), San Francisco, CA, USA, IEEE, 2013. 1373-1376.

[9] Meiru Che and Dewayne E. Perry. 2011. Scenario-based architectural design decisions documentation and evolution. In Proceedings of Engineering of Computer Based Systems (ECBS'11), Las Vegas, NV, (27-29 April 2011), IEEE, 2011, 216-225.

[10] Meiru Che and Dewayne E. Perry. 2012. Managing architectural design decisions documentation and evolution. In Proceedings of 6th International Journal of Computers, 2012, 137-148.

[11] M. Shahin, P. Liang and M.R. Khayyambashi. 2009. Architectural design decision: Existing models and tools. In Proceedings of Software Architecture, 2009 & European Conference on Software Architecture. WICSA/ECSA 2009. Joint Working IEEE/IFIP Conference, IEEE, 2009, 293-296.

[12] M. Shahin, P. Liang, and M.R. Khayyambashi. 2009. A Survey of Architectural Design Decision Models and Tools. Technical Report SBU-RUG-2009-SL-01. http://www.cs.rug.nl/search/uploads/Publications/shahin2009sad.pdf, visited on 8 July 2014.

7. AUTHORS BIOGRAPHIES

Hind Alamin Mohamed BSIT and MSCS, is a lecturer in Software Engineering department, College of Computer Science and Information Technology at Sudan University of Science and Technology (SUST). She has participated in the Scientific Forum for Engineering and Computer Students (December 2005) in SUDAN, and had the first prize of the Innovation and Scientific Excellence for the best graduated project on computer science in 2005. She has been teaching in the areas of Software Engineering and Computer Science since 2006. In 2010 she was the head of Software Engineering Department till December 2012. She is currently a PhD candidate in Software Engineering since 2013.

Hany H. Ammar BSEE, BSPhysics, MSEE, and PhD EE, is a Professor of Computer Engineering in the Lane Computer Science and Electrical Engineering department at West Virginia University. He has published over 170 articles in prestigious international journals and conference proceedings. He is currently the Editor in Chief of the Communications of the Arab Computer Society On-Line Magazine. He is serving and has served as the Lead Principal Investigator in the projects funded by the Qatar National Research Fund under the National Priorities Research Program. In 2010 he was awarded a Fulbright Specialist Scholar Award in Information Technology funded by the US State Department - Bureau of Education and Cultural Affairs. He has been the Principal Investigator on a number of research projects on Software Risk Assessment and Software Architecture Metrics funded by NASA and NSF, and projects on Automated Identification Systems funded by NIJ and NSF. He has been teaching in the areas of Software Engineering and Computer Architecture since 1987. In 2004, he co-authored a book entitled Pattern-Oriented Analysis and Design: Composing Patterns to Design Software Systems, Addison-Wesley. In 2006, he co-authored a book entitled Software Engineering: Technical, Organizational and Economic Aspects, an Arabic Textbook

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 791 - 794, 2014, ISSN:- 2319–8656

Dynamic Resource Provisioning with Authentication in Distributed Database

Anju Aravind K
Shree Venkateshwara Hi-Tech Engineering College
Gobi, India

Dr. T. Senthil Prakash
Shree Venkateshwara Hi-Tech Engineering College
Gobi, India

M. Rajesh
Shree Venkateshwara Hi-Tech Engineering College
Gobi, India

Abstract: Data center have the largest consumption amounts of energy in sharing the power. The public cloud workloads of different priorities and performance requirements of various applications [4]. Cloud data center have capable of sensing an opportunity to present different programs. In my proposed construction and the name of the security level of imperturbable privacy leakage rarely distributed cloud system to deal with the persistent characteristics there is a substantial increases and information that can be used to augment the profit, retrenchment overhead or both. Data Mining Analysis of data from different perspectives and summarizing it into useful information is a process. Three empirical algorithms have been proposed assignments estimate the ratios are dissected theoretically and compared using real Internet latency data recital of testing methods.

Keywords: Mining, MD5, green computing, workload imitation, power consumption

1. INTRODUCTION

Most data centers today have a three or four tier hierarchical networking structure. Three tier network architectures were designed around client-server applications and single-purpose of application server. Client and server applications are caused traffic to flow primarily in patterns: from a server up to the data center core, to the environment core where it moves out to the internet. These large core switches usually contain the vast majority of the intelligence in the network. The cost of builds and operates the large computer platform and a focus on service quality and cost-efficient driving will require cost estimation and on the capacity of processing and storage.

The cloud consists of
Private Cloud: The infrastructure is provisioned for exclusive use by a single organization.
Public Cloud: The infrastructure is provisioned for open use by the general public.
Hybrid Cloud: There is a system infrastructure of two or more distinct cloud Infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability.
Community Cloud: Consumers who share the concerns of the infrastructure for personal use by the provision of a particular community.

While Cloud computing is not equivalent to virtualization, virtualization technology is heavily used to operate a cloud environment system. In the virtualization have one of the host that ran a single operating system now has the ability to run multiple guest operating systems as virtual machines [VMs]. The VMs are created for fast and easily data storage in a cloud environment. The infrastructure environment is invisible for abstracted from the consumer. The computing of firmware system to provide virtual machine and it allows to operating directly on underlying hardware but within specified constraint. It is the software that manages communications between the physical server memory and CPU or processing capability and the virtual machines that are running. The software allows VMs to be quickly provisioned or decommissioned. So the data center contains delay process, security process, mining process and also cost efficiency.

First, in data centre have different perspectives and useful information of data mining [DM]. In DM information that can be used to increase the revenue and it is the process of summarizing the costs or both are reduced. It allows the user to analyze the data with different dimensions or angles. To categorize and summarize the relationships are identified the data. The proposed method of Support Vector Machine for data mining system [DMS] is designed to take the advantage of powerful processors and shared pools to the calculation is performed using the message passing paradigm for data that is distributed across processors. The calculation results are then collected, and the process is repeated with new data processor.

Second, data center security concerns to secure all aspects of cloud data [2]. Many of these features are not unique in cloud system: Irrespective of the data stored on it is vulnerable attack. Therefore the cloud computing have security, access control, malware protection structures, reducing attack surfaces, safety design and implementation of all data including computing security. In this proposed method having MD5 is used to secure the data in data center.

Third, resource provisioning in cloud computing over the last few years, has emerged as new computing model to allowing utility based delivery of services to end users[1]. Cloud computing relies on virtualization technologies to provide on-demand resources according to the end user needs but problems with resource allocation pose an issue for the management of several data centers. In my proposed gossip protocol of green computing based virtualization can be used to increase energy efficiency in substrate networks, allowing consolidation through the virtual hosting of different resources on the same substrate resource. Migrating resources virtually allows the network to balance its overall energy load and reduce the total power consumption of the data center. The dynamic provisioning concept is used to allocate the resources dynamically.

2. RELATED WORK

2.1 Datacenter

The datacenter is the collection of servers, the cloud computing and information technology between physical servers to migrate the cloud computing services, virtualized data center, emerging paradigm changes and executives of the largest independent provider [7]. VM virtualizations and migration capabilities to integrate their computer services and the minimum number of physical servers used to process the data center such as mining processing, security purpose, load balancing, server establishment, online maintenance, proactive fault tolerance and VM migration. Information's are use of the cloud computing to provide services for worldwide users. The consumer scientific controls the commercial domains hosting gives pervasive applications have costs and environmental contribution to the carbon footprint of data centers and cloud hosting applications that consume huge amounts of electrical energy. Therefore, reduces the environmental impact of cloud computing with the help of green computing and gossip protocol.

2.2 Database

Database systems serving cloud platforms must serve large numbers of applications. In addition to managing tolerant with small data footprints, different textures and shapes with variable load patterns such data platforms must minimize their operating costs by efficient resource sharing. The persistent database have the files are stored in network attached storage. VM migrate the database cache and the state of active transactions to ensure minimal impact on transaction execution while allowing transactions active during migration to continue executions and also guarantee the concurrency while ensuring correctness during failures [8].

2.3 Resource Allocation

Dynamic Resource Management for Cloud Computing paradigm is an active area of research. The cost varies are considerably depending upon the configuration of resources by using them [6]. Efficient management of resources, cloud providers and users have the prime interests depending upon the size, capacity and flexibility for the cloud have been managing software which is able to use the hardware resources to succeed, and argued that the critical alone to provide the desired performance [5]. The successful resources management in the context of the resource constraints for the best private cloud solution, initial placement and load balancing when the resources to offers rich set. For example, during the peak of banking applications based on customer needs, they have number of servers that can be accessed from the cloud. In a moment have shut down the server and it can be used to save the energy.

3. METHODOLOGY

3.1 Support Vector Machine

The support vector machine (SVM) is a training algorithm for learning classification and regression rules of data [3]. The SVM is used to find the best classification functions to distinguish between members of the two classes in the trained data. The metrics for the best classification function can be realized geometrically and a linear classification function corresponds to a separating hyperplane $f(x)$ that passes through the middle and separating of the two classes. A function is determined the new data instance of x_n can be classified by simply testing the sign of the function $f(x_n)$: x_n belongs to the positive class if $f(x_n) > 0$.

3.2 MD5 Method

Step 1 – Append padded bits:
– The length of the messages 448 modulo of 512 that is similar to the modular padded.

• 64 bits to 512 bits long means they are just shy of being extended.

– The value of 1 bit is appended to the message and then the value 0 bit also appended so that the length in bits are equals 448 modulo 512.

Step 2 – Append length:

– 64 bit representation of b is appended to the result of the previous step.

– Right that has a length of 512 bits of message.

Step 3 –Initialize MD Buffer

• A four-word buffer (A, B, C, D) is used to compute the message digest.

– Here every one has A, B, C, D is a 32-bit register

- Word A: 01 23 45 67 etc., the following hexadecimal values of these registers are initialized.

Step 4 –The 16 word blocks for process message.

– The input of three 32-bit words and getting the output of one 32-bit word four sub-functions.

F(X,Y,Z) = XY v not(X) Z

The bits X,Y and Z are independent and also unbiased. The each bit of F(X,Y,Z), G(X,Y,Z), H(X,Y,Z), and I(X,Y,Z) will be independent and unbiased.

Step 5 –output

– Output A,B,C,D of the message digest.

– The ouput is start with the low order byte A and end with the high-order byte D.

3.3 GREEN AND GOSSIP

To initiate more environmental - friendly computing practices. There are some steps take toward a green computing strategy. Green Resource Allocator: It is act as the interface between

the Cloud infrastructure and consumers. The interactions of the following components to support energy efficient resource management. The components are Arbitrator, Overhaul-Prediction, Client Focused, Budget Processing, Power Consuming, Overhaul-Router, Storage Leadership and Gauge.

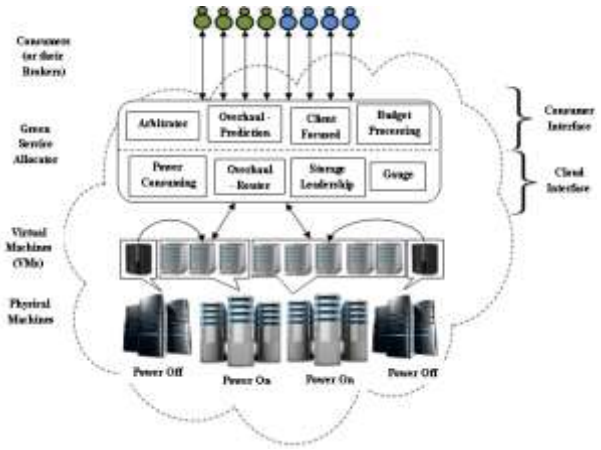


Figure: 3.1 Green Cloud Computing Architecture

Gossip huge clouds of green computing resource allocation based they aim to reduce server power consumption by integration with the specific objectives are sensible the resource allocation protocol and propose a common rumors. Under load the ethical and fair allocation of CPU resources to clients. The simulation result and the key performance of metrics for resource allocation process and suggest that it is the body has do not change with increasing size.

4. SYSTEM IMPLEMENTATION

4.1 General

In general, user interface design for consumer requests and getting response from the server. The consumer is valid provider means they are getting the further processes. It is used to interact with client and server for request and responding from the cloud storage.

4.2 Implementation of Physical Machine

In this module, to implement n number of physical server and which is interconnected with data storage. Each and every physical server have been separate identification like server IP, server port and server instance name. in this all the data stored in the storage device which the data can be applicable to source through physical server.

4.3 Cloud Environment mining setup system

It is provide the dynamic allocation of consumer request to the particular physical services and the physical server retrieve the consumers request based information from the storage device. The response of physical server is applicable for available data only. Cloud environment use the concept of SVM to populate the data from data centers.

4.4 Highly Secure Client Processing

In this system using MD5 based concept of achieving the securable data transmission between the consumer and the server. This MD5 is converted into the data normal format to undefined format. In this application apply the highly security for data transmission client request processing physical machine data population server mining user identification by server using the request name.

4.5 Efficient Server Provisioning

Dynamic allocation has been user request to the physical server in done by cloud environment by using the concept of protocol named as gossip [5]. This protocol is sufficient protocol for dynamic resource allocation and it gives response to the client at exact query matching server provisioning approach for minimized the cost, reduce the time and quick response. In order to assemble the cloud environmental setup and physical server storage device is very expensive but they are applying the mining setup. Show that it is must like expensive.

5. EXPERIMENTAL RESULT

The implementation of my concept is to create the n number of virtual machines and physical machines. In this machines have n number of information's are stored. This physical machine contains java connection based classes and service based concepts. In cloud environment system distributed set up of mined servers. The cloud server's shows all the information of the physical server in the data center.

In data center having mining setup for retrieve the data from the data storage. In data storage wants to store the number of files with the help of query processing from server. In this method using SVM for classify the data for the user query searched from the server.

The data center is the large storage network. The network wants to secure the stored information for the storage devices by using the cryptography technique. In this concept am using MD5 method for creating number of keys to secure the data from the storage devices. The valid user's are only views the information about the user queries. The key based concept is achieving the more secure for storage devices.

Finally, the server client communication is the very large process. In this having number of resources and it also intermediate process for file transfers to the user's and data storage. It is the nonstop processes so here using gossip protocol for green computing process. It is an automatically allocate the resources for file sharing in the data center.

6. CONCLUSION

To conclude that the concept have so many process in the cloud environment. The cloud environment has number of virtual machines and physical machines. These machines are used to store the number of data in the storage devices. The storage devices have been number of processes to mining the data.

In this the data can be find out from the storage device. The data retrieval by the user's from the cloud storage devices. In this the user given the request and get the response from the server. The SVM method is used for mined the data from the storage device. This method is used to classify the data and it is very efficient to gather some information without any unknown data.

The second process is to secure the data by using the method of MD5. The cryptographic technique is used to secure the information storage by using the key values. So the information storage is getting more secure without any leakage in the cloud environment.

The third process to find the delay provisioning for data center. The data center having so many request and response

process for user's from the storage devices. In this device having continue processes so the machines want to overcome the delay processing by using the method of gossip protocol for green computing.

7. REFERENCES

[1] Qi Zhang, Mohamed Faten Zhani, Raouf Boutaba, Joseph L. Hellerstein "Dynamic Heterogeneity Aware Resource Provisioning in the Cloud" Proc.IEEE Trans. Cloud computing,2014.

[2] Jun Zhou, Xiaodong Lin, Xiaolei Dong, Zhenfu Cao "PSMPA: Patient Self-controllable and Multi-level Privacy-preserving Cooperative Authentication in Distributed m-Healthcare Cloud Computing System" Proc. IEEE Tans. Parallel and Distributed System, 2014.

[3] Lu Zhang and Xueyan Tang "The Client Assignment Problem for Continuous Distributed Interactive Applications: Analysis, Algorithms, and Evaluation" Proc. IEEE Trans. Parallel and Distributed System, 2014.

[4] Q. Zhang, M.F. Zhani, R. Boutaba, and J.L. Hellerstein, "HARMONY: Dynamic Heterogeneity-Aware Resource Provisioning in the Cloud,"Proc. IEEE Int'l Conf. Distributed Computing Systems (ICDCS), 2013.

[5] Q. Zhang, M.F. Zhani, Q. Zhu, S. Zhang, R. Boutaba, and J.L. Hellerstein, "Dynamic Energy-Aware Capacity Provisioning for Cloud Computing Environments," Proc. ACM Int'l Conf. Autonomic Computing (ICAC), 2012.

[6] Lu Zhang, Xueyan Tang, "Optimizing Client Assignment for Enhancing Interactivity in Distributed Interactive Applications," Proc. IEEE/ACM Transaction on Networking, 2012.

[7] P. Morillo, J. Orduna, M. Fernandez, and J. Duato, "Improving the performance of distributed virtual environment systems," IEEE Trans. Parallel Distrib. Syst., vol. 16, no. 7, pp. 637–649, Jul. 2005.

[8] J. Sun and Y. Fang, Cross-domain Data Sharing in Distributed Electronic Health Record System, IEEE Transactions on Parallel and Distributed Systems, vol. 21, No. 6, 2010.

Authors

MS. Anju Aravind K, received the Bachelor of Engineering in Anna University, TamilNadu, India in 2011. PG Scholar Currently persuing her M.E CSE degree in shree Venkateshwara Hi-Tech Engg College, Gobi, TamilNadu, India.

Dr.T.Senthil Prakash received the Ph.D. degree from the PRIST University, Thanjavur, India in 2013 and M.E(CSE) degree from Vinayaka Mission's University, Salem, India in 2007 and M.Phil.,MCA.,B.Sc(CS) degrees from Bharathiyar University, Coimbatore India, in 2000,2003 and 2006 respectively, all in Computer Science and Engineering. He is a Member in ISTE New Delhi, India, IAENG, Hong Kong..IACSIT, Singapore SDIWC, USA. He has the experience in Teaching of 10+Years and in Industry 2 Years. Now He is currently working as a Professor and Head of the Department of Computer Science and Engineering in Shree Venkateshwara Hi-Tech Engineering College, Gobi, Tamil Nadu, and India. His research interests include Data Mining, Data Bases, Artificial Intelligence, Software Engineering etc.,He has published several papers in 17 International Journals, 43 International and National Conferences.

Mr.M.Rajesh, Received the Bachelor of Engineering in Anna University, TamilNadu, India in 2007 and Master of Engineering from Kongu Engineering College of India in 2012. Currently he is doing Ph.D at Bharath University, Chennai. His research interests include cloud computing in resource provisioning.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 795 - 798, 2014, ISSN:- 2319–8656

Guarding Against Large-Scale Scrabble In Social Network

Geerthidevi K G
Shree Venkateshwara Hi-Tech Engineering College
Gobi, India

Dr. T. Senthil Prakash
Shree Venkateshwara Hi-Tech Engineering College
Gobi, India

Prakadeswaran M.E
Shree Venkateshwara Hi-Tech Engineering College
Gobi, India

Abstract: Generally, the botnet is one of the most dangerous threats in the network. It has number attackers in the network. The attacker consists of DDOS attack, remote attack, etc., Bots perform perform repetitive tasks automatically or on a schedule over the internet, tasks that would be too mundane or time-consuming for an actual person. But the botnets have stealthy behavior as they are very difficult to identify. These botnets have to be identified and the internet have to be protected. Also the the activity of botnets must be prevented to provide the users, a reliable service. The past of botnet detection has a transaction process which is not secure. A efficient stastical data classifier is required to train the botent preventions system. To provide the above features clustering based analysis is done. our approach can detect and profile various P2P applications rather than identifying a specific P2P application. Anomaly based detection technique is used to obtain this goal.

Keywords: Botnet, anomaly base detection, hash function, DDOS

1. INTRODUCTION

A botnet is a collection of Internet-connected programs communicating with other similar programs in order to perform tasks. Botnets sometimes compromise computers whose security defenses have been breached and control conceded to a third party. [1]It is remotely controlled by an attacker through a command and control (C&C) channel. Botnets serve as the infrastructures responsible for a variety of cyber-crimes, such as spamming, distributed denial of-service (DDoS) attacks, identity theft, click fraud, etc. The C&C channel is an essential component of a botnet because botmasters rely on the C&C channel to issue commands to their bots and receive information from the compromised machines.

Bots perform perform repetitive tasks automatically or on a schedule over the internet, tasks that would be too mundane or time-consuming for an actual person. Search engines use them to surf the web and methodically catalogue information from websites, trading sites make them look for the best bargains in seconds, and some websites and services employ them to deliver important information like weather conditions, news and sports, currency exchange rates.

Unfortunately, not all bots roaming the internet are useful and harmless. Cyber crooks have also noticed their potential and have come up with malicious bots – programs designed to secretly install themselves on unprotected or vulnerable computers and carry out whatever actions they demand. And that could be anything from sending spam to participating in a distributed denial of service attack (DDoS) that brings down entire websites

Once infected, your computer becomes part of a botnet – a network of infected or zombie-computers controlled from the distance by a cybercriminal who rented it to carry out his illegal plans. So not only is your computer infected and your internet security compromised, but your system resources and your bandwidth are rented out to the highest bidder to help them attack other unsuspecting users or even legitimate businesses. This huge potential for cybercrime makes these botnets what some security experts believe to be the most dangerous threat on the internet today.

Such networks comprising hundreds or thousands of infected devices have the resources needed to perform high-scale malicious actions such as: (1) Mass-spam delivery that floods millions of inboxes in a matter of seconds (2) DoS and DDoS attacks that crash entire websites and can put legitimate businesses in serious trouble (3) Brute-force hacking attacks by cracking passwords and other internet security measures (4) Identity theft and internet fraud by collecting private information from infected users

Bots can sneak up on you in many ways. They can use the vulnerabilities and outdated software in your system to infect it while you're casually surfing the web. They can be delivered by Trojans or questionable software you get tricked into downloading (like rogue antivirus programs). Or they can be sent directly to your inbox as an email attachment by spammers.

Botnets perform many malicious activity in internet like sending spams to emails, increasing network traffic and even takes control of the system by running Trojans. But the botnets have stealthy behavior as they are very difficult to identify. These botnets have to be identified and the internet have to be protected. The information shared in social media are sensitive and personal. Hence the activity of botnets must be prevented to provide the users, a reliable service.

To provide the above features clustering based analysis is done. our approach can detect and profile various P2P applications rather than identifying a specific P2P application. Anomaly based detection technique is used to obtain this goal.

2. RELATED WORKS

Many approaches have been proposed to detect botnets have been proposed. For example, BotMiner [7] identifies a group of hosts as bots belonging to the same botnet if they share similar communication patterns and meanwhile perform similar malicious activities, such as scanning, spamming,

exploiting, etc.[4] Unfortunately, the malicious activities may be stealthy and non-observable A efficient stastical data classifier is required to train the botent preventions system. Acquiring such information is a challenging task, thereby drastically limiting the practical use of these methods. Some of the older approach involves content signature, encryptions, profiling, fixed source port. our approach does not need any content signature. our analysis approach can estimate the active time of a P2P application, which is critical for botnet detection

3. SYSTEM DESIGN

A Botmaster has to be designed with P2P protocol. Therefore P2P bots exhibit some network traffic patterns that are common to other P2P client applications either legitimate or malicious. Hence our system is divided into two phases. In the first phase, we aim at detecting all hosts within the monitored network that engage in P2P communications. We analyze raw traffic collected at the edge of the monitored network and apply a pre-filtering step to discard network flows that are unlikely to be generated by P2P[1]. We then analyze the remaining traffic and extract a number of statistical features to identify flows generated by P2P clients. In the second phase, our system analyzes the traffic generated by the P2P clients and classifies them into either *legitimate* P2P clients or P2P *bots*. Specifically, we investigate the active time of a P2P client and identify it as a *candidate* P2P bot if it is persistently active on the underlying host. We further analyze the overlap of peers contacted by two *candidate* P2P bots to finalize detection. After analyzing with the use of anomaly based detection algorithm the network has to be revoked from malwares.

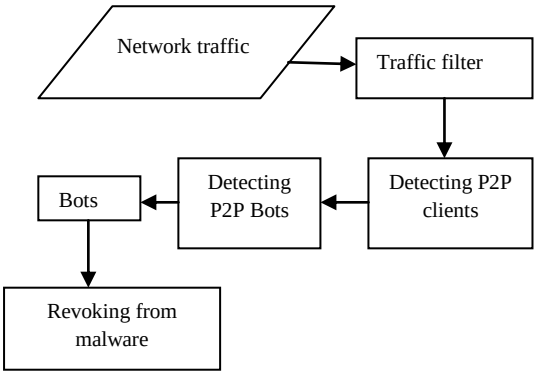


Fig 1: System architecture

3.1 Detecting P2P client

Traffic filter is used to sort out th traffic which is unlikely to P2P networks. In this first phase, fine grained detection of P2P botenets is implemented. This component is responsible for detecting P2P clients by analyzing the remaining network flows after the Traffic Filter component. For each host h within the monitored network we identify two flow sets, denoted as Stcp(h) and Sudp(h), which contain the flows related to successful outgoing TCP and UDP[6] connection, respectively.

To identify flows corresponding to P2P control messages,we first apply a flow clustering process intended to group together similar flows for each candidate P2P node *h*. Given sets of flows *Stcp(h)* and *Sudp(h)*, we characterize each flow using a vector of statistical features $v(h) = [Pkts, Pktr, Bytes, Byter]$, in which *Pkts* and *Pktr* represent the number of packets sent and received, and *Bytes* and *Byter* represent the number of bytes sent and received, respectively.

The distance between two flows is subsequently defined as the *euclidean distance* of their two corresponding vectors. We then apply a clustering algorithm to partition the set of flows into a number of clusters. Each of the obtained clusters of flows, *Cj (h)*, represents a group of flows with similar size.

Flows corresponding to ping/pong and peer-discovery share similar sizes, and hence they are grouped into two clusters (FC1 and FC2), respectively. Since the number of destination BGP prefixes involved in each cluster is larger, we take FC1 and FC2 as its fingerprint clusters. A fingerprint cluster summary, (Pkts , Pktr , Bytes , Byter , proto), represents the protocol and the average number of sent/received packets/bytes for all the flows in this fingerprint cluster. We implemented the flow analysis component and identified fingerprint cluster for the sample P2P traces including two traces.

3.2 Detecting P2P bots

To detect the bots coarse grained detection method is used. Since bots are malicious programs used to perform profitable malicious activities, they represent valuable assets for the botmaster, who will intuitively try to maximize utilization of bots. This is particularly true for P2P bots[5] because in order to have a functional overlay network (the botnet), a sufficient number of peers needs to be always online. In other words, the active time of a bot should be comparable with the active time of the underlying compromised system.

The distance between each pair of hosts is computed. We apply hierarchical clustering, and group together hosts according to the distance defined above. In practice the hierarchical clustering algorithm will produce a dendrogram (a tree-like data structure). The dendrogram expresses the "relationship" between hosts. The closer two hosts are, the lower they are connected at in the dendrogram. Two P2P bots in the same botnet should have small distance and thus are connected at lower level. In contrast, legitimate P2P applications tend to have large distances and consequently are connected at the upper level. We then classify hosts in dense clusters as P2P bots, and discard all other clusters and the related hosts, which we classify as legitimate P2P clients.

4. SYSTEM IMPLEMENTATION

Out of four components in our system, "Traffic Filter" and "Coarse-Grained Detection of P2P Bots" have linear complexity since they need to scan flows only once to identify flows with destination addresses resolved from DNS queries or calculate the active time. Other two components, "Fine-Grained Detection of P2P Clients" and "Fine-Grained P2P Detection of P2P Bots", require pairwise comparison for distance calculation

We use a two-step clustering approach to reduce the time complexity of "Fine-Grained P2P Client Detection". For the first-step clustering, we use an efficient clustering algorithm to aggregate network flows into K sub-clusters, and each subcluster contains flows that are very similar to each other. For the second-step clustering, we investigate the global

distribution of sub-clusters and further group similar sub-clusters into clusters.

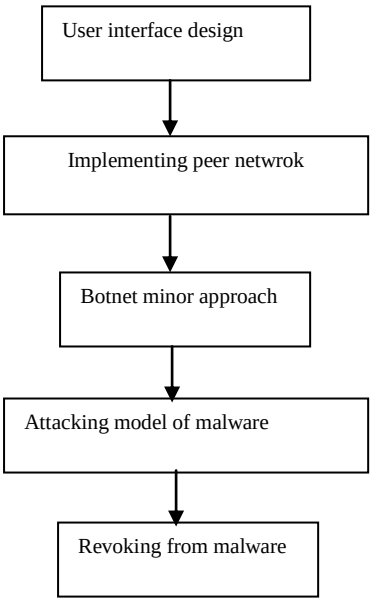
The distance of two flows is defined as the Euclidean distance of their corresponding vectors, where each vector [Pkts , Pktr , Bytes , Byter] represents the number of packets/ bytes that are sent/received in a flow.

For the second-step clustering, we use hierarchical clustering with DaviesBouldin validation[8] to group sub-clusters into clusters. Each sub-cluster is represented using a vector ([Pkts , Pktr , Bytes , Byter]), which is essentially the average for all flow vectors in this sub-cluster.

Hierarchical clustering is used to build a dendrogram. Finally, DaviesBouldin validation is employed to assess the global distribution of inter- and intra-cluster distances of clusters based on various clustering decisions and yield the best cut for the dendrogram. The two-step clustering algorithm has the time complexity of O(nK I + K2).

4.1 Modules

The goal of guarding the large scale scrabble in social network is implemented by the following modules,



4.1.1 User Interface Design

The user interaction is effective operation and control of the machine on the user‘s. The user interface module has login and registration phases. The registration phase gets details from user and stores it in database. It also checks the details are valid or not.

4.1.2 Implementing peer network

The peer network contain decentralized networks. All the nodes contains separate IP address and separate port number. The peer one node have stored separate list of files which in the global respository.

4.1.3 Botnet minor approach

The global respository contains the decentralized network details. The botnet minor store and reteive the information about port and IP details from the database. Identification scenario always visible botnet minor. If any dispute in the identification scenario overall network may be crashed.

4.1.4 Attacking model of Malware

Botnet minor contain all the details about the peer network. The botnet minor handles all the request processed by the decentralized network. The botnet major attack decentralized scenario spread the warm data to the peer network. The node connected with the attacked node that specific node also get the warm data.

4.1.5 Revoking the network from Malware

Data matching have the law data and the original data. The proposed technical approach can identify the warm data it is spreaded by the botnet. Revoke the original data instead of warm data it can identify the problem and revoke the botnet minor from the attacking model.

5. EXPERIMENTAL RESULTS

We prepared a data set (D) for evaluation. Specifically, we randomly selected half [8] of the P2P bots from NETbots .Then for each of the 5 P2P applications we ran, we randomly selected one out of its two traces from NETP2P and overlaid its traffic to the traffic of a randomly selected host We applied our detection system on data set D. The traffic filter drastically reduced the workload for the whole system. As indicated in Figure 4, it reduced the number of hosts subject to analysis by 67% (from 953 to 316) but retained all P2P clients.

Among 26 P2P clients identified in the previous step, 25 out of them exhibit persistent P2P behaviors. We further evaluate the similarity of fingerprint clusters and peer IPs for each pair of persistent P2P clients and
derive a dendrogram.

If botmasters get to know about our detection algorithm, they could attempt to modify other bots’ network behavior to evade detection. This situation is similar to evasion attacks against other intrusion detection systems

6. CONCLUSION

To summarize, although our system greatly enhances and complements the capabilities of existing P2P botnet detection systems, it is not perfect. We should definitely strive to develop more robust defense techniques, where the aforementioned discussion outlines the potential improvements of our system.

In this paper, we presented a novel botnet detection system that is able to identify *stealthy* P2P botnets, whose malicious activities may not be observable. To accomplish this task, we derive *statistical fingerprint*s of the P2P communications to first detect P2P clients and further distinguish between those that are part of legitimate P2P networks (e.g., filesharing networks) and P2P bots. We also identify the performance bottleneck of our system and optimize its scalability. The evaluation results demonstrated that the proposed system accomplishes high accuracy on detecting stealthy P2P bots and great scalability.

7. REFERENCES

[1] S. Stover, D. Dittrich, J. Hernandez, and S. Dietrich, "Analysis of the storm and nugache trojans: P2P is here," in Proc. USENIX, vol. 32. 2007, pp. 18–27.

[2] Junjie Zhang, Roberto Perdisci, Wenke Lee, Xiapu Luo, and Unum Sarfraz, "Building a Scalable System for Stealthy P2P-Botnet Detection", IEEE transactions on information forensics and security, vol. 9, no. 1, january 2014

[3] Pratik Narang, Subhajit Ray, Chittaranjan Hota, Venkat Venkatakrishnan, "PeerShark: Detecting Peer-to-Peer Botnets by Tracking Conversations",2014 IEEE Security and Privacy Workshops

[4] P. Porras, H. Saidi, and V. Yegneswaran, "A multi-perspective analysis of the storm (peacomm) worm," Comput. Sci. Lab., SRI Int., Menlo Park, CA, USA, Tech. Rep., 2007.

[5] P. Porras, H. Saidi, and V. Yegneswaran. (2009). Conficker C Analysis [Online]. Available: http://mtc.sri.com/Conficker/addendumC/index.html

[6] G. Sinclair, C. Nunnery, and B. B. Kang, "The waledac protocol: The how and why," in Proc. 4th Int. Conf. Malicious Unwanted Softw., Oct. 2009, pp. 69–77.

[7] G. Gu, R. Perdisci, J. Zhang, and W. Lee, "Botminer: Clustering analysis of network traffic for protocol- and structure-independent botnet detection," in Proc. USENIX Security, 2008, pp. 139–154.

[5] R. Lemos. (2006). Bot Software Looks to Improve Peerage[Online]Available:http://www.securityfocus.com/news/11390

[6] Y. Zhao, Y. Xie, F. Yu, Q. Ke, and Y. Yu, "Botgraph: Large scale spamming botnet detection," in Proc. 6th USENIX NSDI, 2009,
pp. 1–14.

[8] T.-F. Yen and M. K. Reiter, "Are your hosts trading or plotting? Telling P2P file-sharing and bots apart," in Proc. ICDCS, Jun. 2010, pp. 241–252.

Authors:

Ms. Geerthidevi K G, PG Scholar Currently persuing her M.E CSE degree in Shree Venkateshwara Hi-Tech Engg College, Gobi, Tamilnadu , India. Her research interests include Networking, Network Security etc.,

Dr.T.Senthil Prakash received the Ph.D. degree from the PRIST University, Thanjavur, India in 2013 and M.E(CSE) degree from Vinayaka Mission's University, Salem, India in 2007 and M.Phil.,MCA.,B.Sc(CS) degrees from Bharathiyar University, Coimbatore India, in 2000,2003 and 2006 respectively, all in Computer Science and Engineering. He is a Member in ISTE New Delhi, India, IAENG, Hong Kong..IACSIT, Singapore SDIWC, USA. He has the experience in Teaching of 10+Years and in Industry 2 Years. Now He is currently working as a Professor and Head of the Department of Computer Science and Engineering in Shree Venkateshwara Hi-Tech Engineering College, Gobi, Tamil Nadu, and India. His research interests include Data Mining, Data Bases, Artificial Intelligence, Software Engineering etc.,He has published several papers in 17 International Journals, 43 International and National Conferences.

Mr.S.Prakadeswaran, received the Bachelor of Engineering in Anna University,Chennai,Tamilnadu in 2008.He received the Master of Engineering in Anna University, Chennai, Tamilnadu in 2013. He has the experience in Teaching of 6+Years. He is currently working as Assistant professor in Shree Venkateshwara Hi Tech Engineering College, Gobichettipalayam, Tamilnadu. His research interest includes Wireless Networks and Pervasive computing. He has published several papers in 4 International Journals

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 799 - 808, 2014, ISSN:- 2319–8656

A Secure, Scalable, Flexible and Fine-Grained Access Control Using Hierarchical Attribute-Set-Based Encryption (HASBE) in Cloud Computing

Prashant A. Kadam
Department of Computer Engineering
JSPM Narhe Technical Campus
Pune, India

Avinash S. Devare
Department of Computer Engineering
JSPM Narhe Technical Campus
Pune, India

Abstract: Cloud Computing is going to be very popular technology in IT enterprises. For any enterprise the data stored is very huge and invaluable. Since all tasks are performed through network it has become vital to have the secured use of legitimate data. In cloud computing the most important matter of concern are data security and privacy along with flexibility, scalability and fine grained access control of data being the other requirements to be maintained by cloud systems Access control is one of the prominent research topics and hence various schemes have been proposed and implemented. But most of them do not provide flexibility, scalability and fine grained access control of the data on the cloud. In order to address the issues of flexibility, scalability and fine grained access control of remotely stored data on cloud we have proposed the hierarchical attribute set-based encryption (HASBE) which is the extension of attribute- set-based encryption(ASBE) with a hierarchical structure of users. The proposed scheme achieves scalability by handling the authority to appropriate entity in the hierarchical structure, inherits flexibility by allowing easy transfer and access to the data in case of location switch. It provides fine grained access control of data by showing only the requested and authorized details to the user thus improving the performance of the system. In addition, it provides efficient user revocation within expiration time, request to view extra-attributes and privacy in the intra-level hierarchy is achieved. Thus the scheme is implemented to show that is efficient in access control of data as well as security of data stored on cloud with comprehensive experiments.

Keywords: Fine-grained access control, attribute-set-based encryption, hierarchical attribute-set-based encryption.

1. INTRODUCTION

Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the data centers that provide those services. The services themselves have long been referred to as Software as a Service (SaaS).The datacenter hardware and software is what we will call a Cloud.

Cloud computing is a web-based application that provides computation, software, infrastructure, platform, devices and other resources to users on the basis of pay as you use. Clients can use cloud services without any installation and the data uploaded on cloud is accessible from anywhere in the world, the only requirement is the computer with active internet connection. As a customizable computing resources and a huge amount of storage space are provided by internet based online services, the shift to online storage has contributed greatly in eliminating the overhead of local machines in storage and maintenance of data. Cloud provides a number of benefits like flexibility, disaster management and recovery, pay-per-use and easy to access and use model which contribute to the reason of switching to cloud. Cloud gives the provision for storage of important data of users. Thus cloud helps to free up the space on the local disk.

Cloud computing has emerged as one of the most influential paradigms in the IT industry. Almost all companies, organizations store their valuable data on the cloud and access it. Due to this security to the cloud is a major concern. Also flexibility, scalability of data stored on cloud which are the system performance parameters and which degrade the system response time required to be handled by cloud systems. They should provide a secure environment and maintenance of data in hierarchy.

The prominent security concern is data storage security and privacy in cloud computing due to its Internet-based data storage and management. The data security issue becomes vital when the data is a confidential data. In cloud computing, users have to give up their data to the cloud service provider for storage and business operations, while the cloud service provider is usually a commercial enterprise which cannot be totally trusted. So the data integrity and privacy of data is at risk.

Flexible and fine-grained access control is strongly desired in the service-oriented cloud computing model. Various schemes which provide access control models have been proposed. But the problem related with these schemes is that they are limited to data owners and service providers which exist in the same trusted domain.

2. EXISTING SYSTEM

2.1 Vipul et al.“ (Abe)Attribute based encryption”. [1]

As more sensitive data is shared and stored by third-party sites on the Internet, there will be a need to encrypt data stored at these sites. One drawback of encrypting data, is that it can be selectively shared only at a coarse-grained level (i.e., giving another party your private key). We develop a new cryptosystem for fine-grained sharing of encrypted data that we call Key-Policy Attribute-Based Encryption (KP-ABE). In our cryptosystem, cipher texts are labeled with sets of attributes and private keys are associated with access structures that control which cipher texts a user is able to decrypt. We demonstrate the applicability of our construction to sharing of audit-log information and broadcast encryption. Our construction supports delegation of private keys which subsumes Hierarchical Identity-Based Encryption (HIBE).

2.2 Rakesh et al. “Attribute-Sets: A Practically Motivated Enhancement to Attribute-Based Encryption”, University of Illinois at Urbana-Champaign, July 27, 2009. [2]

In distributed systems users need to share sensitive objects with others base on the recipients’ ability to satisfy a policy. Attribute-Based Encryption (ABE) is a new paradigm where such policies are specified and cryptographically enforced in the encryption algorithm itself. Cipher text-Policy ABE (CP-ABE) is a form of ABE where policies are associated with encrypted data and attributes are associated with keys. In this work we focus on improving the flexibility of representing user attributes in keys. Specifically, we propose Cipher text Policy Attribute Set Based Encryption (CP-ASBE) - a new form of CP-ABE - which, unlike existing CP-ABE schemes that represent user attributes as a monolithic set in keys, organizes user attributes into a recursive set based structure and allows users to impose dynamic constraints on how those attributes may be combined to satisfy a policy. We show that the proposed scheme is more versatile and supports many practical scenarios more naturally and efficiently. We provide a prototype implementation of our scheme and evaluate its performance overhead

2.3 Pankaj et al. “Cloud Computing Security Issues in Infrastructure as a Service”, 2012. [3]

Cloud computing is current buzzword in the market. It is paradigm in which the resources can be leveraged on peruse basis thus reducing the cost and complexity of service providers. Cloud computing promises to cut operational and capital costs and more importantly it let IT departments focus on strategic projects instead of keeping datacenters running. It is much more than simple internet. It is a construct that allows user to access applications that actually reside at location other than user’s own computer or other Internet-connected devices. There are numerous benefits of this construct. For instance other company hosts user application. This implies that they handle cost of servers, they manage software updates and depending on the contract user pays less i.e. for the service only. Confidentiality, Integrity, Availability, Authenticity, and Privacy are essential concerns for both Cloud providers and consumers as well. Infrastructure as a Service (IaaS) serves as the foundation layer for the other delivery models, and a lack of security in this layer will certainly affect the other delivery models, i.e., PaaS, and SaaS that are built upon IaaS layer. This paper presents an elaborated study of IaaS components’ security and determines vulnerabilities and countermeasures. Service Level Agreement should be considered very much importance.

2.4 John et al. “(CP-ABE) Cipher text-Policy Attribute-Based Encryption” John et al. [4]

In several distributed systems a user should only be able to access data if a user posses a certain set of cre- dentials or attributes. Currently, the only method for enforcing such policies is to employ a trusted server to store the data and mediate access control. However, if any server storing the data is compromised, then the confidentiality of the data will be compromised. In this paper we present a system for realizing complex access control on encrypted data that we call Cipher text-Policy Attribute-Based Encryption. By using our techniques encrypted data can be kept confidential even if the storage server is not trusted; moreover, our methods are secure against collusion attacks. Previous Attribute- Based Encryption systems used attributes to describe the encrypted data and built

policies into user's keys; while in our system attributes are used to describe a user's credentials, and a party encrypting data deter-mines a policy for who can decrypt. Thus, our meth-ods are conceptually closer to traditional access control methods such as Role-Based Access Control (RBAC). In addition, we provide an implementation of our sys- tem and give performance measurements.

2.5 Ayad et al. "Enabling Data Dynamic and Indirect Mutual Trust for Cloud Computing Storage System", 2012. [6]

In this paper, we propose a cloud-based storage scheme that allows the data owner to benefit from the facilities offered by the CSP and enables indirect mutual trust between them. The proposed scheme has four important features: (i) it allows the owner to outsource sensitive data to a CSP, and perform full block-level dynamic operations on the outsourced data, i.e., block modification, insertion, deletion, and append, (ii) it ensures that authorized users (i.e., those who have the right to access the owner's file) receive the latest version of the outsourced data, (iii) it enables indirect mutual trust between the owner and the CSP, and (iv) it allows the owner to grant or revoke access to the outsourced data. We discuss the security issues of the proposed scheme. Besides, we justify its performance through theoretical analysis and experimental evaluation of storage, communication, and computation overheads.

2.6 Guojun et al. "Hierarchical attribute-based encryption and scalable user revocation for sharing data in cloud servers", 2011. [8]

With rapid development of cloud computing, more and more enterprises will outsource their sensitive data for sharing in a cloud. To keep the shared data confidential against untrusted cloud service providers (CSPs), a natural way is to store only the encrypted data in a cloud. The key problems of this approach include establishing access control for the encrypted data, and revoking the access rights from users when they are no longer authorized to access the encrypted data. This paper aims to solve both problems. First, we propose a hierarchical attribute-based encryption scheme (HABE) by combining a hierarchical identity-based encryption (HIBE) system and a ciphertext-policy attribute-based encryption (CP-ABE) system, so as to provide not only fine-grained access control, but also full delegation and high performance. Then, we propose a scalable revocation scheme by applying proxy re-encryption (PRE) and lazy re-encryption (LRE) to the HABE scheme, so as to efficiently revoke access rights from users.

2.7 Qin et al. "Hierarchical Attribute-Based Encryption for Fine-Grained Access Control in Cloud Storage Services". [9]

Cloud computing, as an emerging computing paradigm, enables users to remotely store their data into a cloud so as to enjoy scalable services on-demand. Especially for small and medium-sized enterprises with limited budgets, they can achieve cost savings and productivity enhancements by using cloud-based services to manage projects, to make collaborations, and the like. However, allowing cloud service providers (CSPs), which are not in the same trusted domains as enterprise users, to take care of confidential data, may raise potential security and privacy issues. To keep the sensitive user data confidential against untrusted CSPs, a natural way is to apply cryptographic approaches, by disclosing decryption keys only to authorized users. However, when enterprise users outsource confidential data for sharing on cloud servers, the adopted encryption system should not only support fine-grained access control, but also provide high performance, full delegation, and scalability, so as to best serve the needs of accessing data anytime and anywhere, delegating within enterprises, and achieving a dynamic set of users. In this paper, we propose a scheme to help enterprises to efficiently share confidential data on cloud servers. We achieve this goal by first combining the hierarchical identity-based encryption (HIBE) system and the cipher text-policy attribute-based encryption (CP-ABE) system, and then making a performance-expressivity tradeoff, finally applying proxy re-encryption and lazy re-encryption to our scheme.

2.8. Patrick et al. "Methods and Limitations of Security Policy Reconciliation". [10]

A security policy is a means by which participant session requirements are specified. However, existing frameworks provide limited facilities for the automated reconciliation of

participant policies. This paper considers the limits and methods of reconciliation in a general-purpose policy model. We identify an algorithm for efficient two-policy reconciliation, and show that, in the worst-case, reconciliation of three or more policies is intractable. Further, we suggest Efficient heuristics for the detection and resolution of intractable reconciliation. Based upon the policy model, we describe the design and implementation of the Ismene policy language. The expressiveness of Ismene, and indirectly of our model, is demonstrated through the representation and exposition of policies supported by existing policy languages. We conclude with brief notes on the integration and enforcement of Ismene policy within the Antigone.

3. PROPOSED SYSTEM

In our propose system instead of showing complete data from cloud we are fetching only those data which is essential for that user. We are not fetching all data so it takes less time for fetching data so system response time is very less due to which system performance increases. We are performing encryption before storing data so even if data get hack by hacker data cannot be easily understand by hacker. We are performing hierarchical structure so even if lower authority is absent for particular days at that time higher authority handle all work of lower authority so work of company will not be stopped. The HASBE scheme for realizing scalable, flexible and fine-grained access control in cloud computing. The HASBE scheme seamlessly incorporates a hierarchical structure of system users by applying a delegation algorithm to ASBE. HASBE not only supports compound attributes due to flexible attribute set combinations, but also achieves efficient user revocation because of multiple value assignments of attributes. We formally proved the security of HASBE based on the security of CP-ABE. Finally, we completed the detailed analysis of proposed scheme, and conducted comprehensive performance analysis and evaluation, which showed its efficiency and advantages over existing schemes.

3.1 Project Scope

1. This system is designed to provide security to data stored on cloud and improve performance of system by showing only the required details requested by an employee.
2. Security is provided by generating a secret key from the various attributes stated in the form which is filled by the employee at the time of registration.
3. This system is designed to provide flexibility of the data where in case of transfer of employee, his data could be transferred to respective location with ease.
4. It also provides scalability in case when an employee is absent his work could be handled by the senior employee securely.

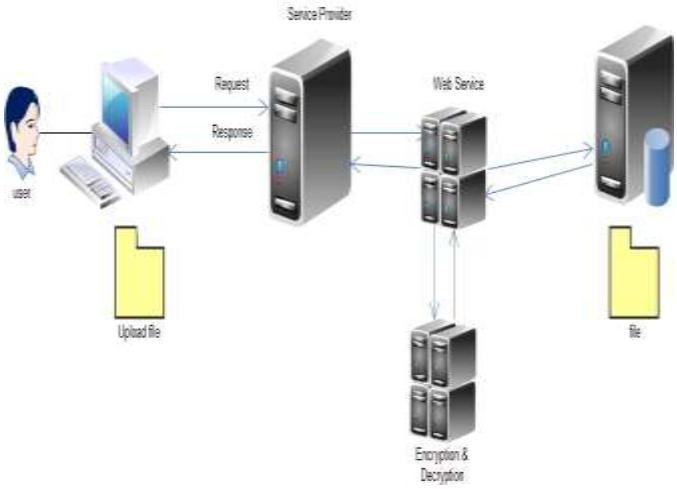


Figure 1. General Architecture of the System

3.1 Methodology

1. Registration and login by user:
In this user fill his/her own complete data. Request is sent to the CEO for confirmation. CEO confirms his/her request and assigns attribute and time period for that user. Once Account get confirm password and key is sent to that user by email so he/she can access his/her account**.**

2. Approve User and Assign attributes:
Out of the selected attributes according the roles defined in hierarchy of the system the attribute visibility access is decided. Each attribute is encrypted.

3. Key Generation and Verification
Key is generated based on the attributes filled by the user in registration form. In attribute key verification, when a key is used for login ,it is first checked with the key stored in the database. If a match is found then user is allowed for further process else the user is rejected for further process.

4. Encryption and decryption of data
User fills his/her data during registration. Once it is click on submit button data is send to encryption algorithm that are RSA and AES. After performing encryption data is stored in encrypted format in database.

5. Access Right:
The user can view the selected attributes of the same level as well as other levels according to the access authority using attribute key.

6. Fine Grained Access
In our propose system instead of showing complete data, the fetching of necessary data is allowed. Due to this system provides a quick response time.

7. Request for extra attribute:

The user can access attributes of same level as inter level counterparts. He can request for extra attributes in case of emergency as well as ease of work.

8. Flexibility
In this module flexibility can be done by suppose user is transfer from one location to another location and for that new location that user's data is not accessible then authority request for accessing data of that user from old location. Once authority got request that data should be access from new location and it is not visible for old location

9. Scalability:
We are performing hierarchical structure so even if lower authority is absent for particular days at that time higher authority handle all work of lower authority so work of company will not be stopped.

10. Efficient User Revocation:
It can be done by two steps request to the admin and response to the user from admin within expiration time.

9. Privacy:
Default it is public but a user can set intra-level privacy by restricting access to attributes.

3.2 Process Summary
Following Processes will be involved in Project:

1. Encrypt data before Insert

After user click on submit button data encrypted using RSA and AESs algorithm. Once data get encrypted it is stored into database and when user wants to retrieve data it again decrypted and shown in original form.

2. Request for New Attributes
In this phase one of the lower authority may absence then at that time higher authority may handle both attributes, one is its own attributes and another is attributes of the lower authority who is absent for particular time period. User can also request for new attribute if needed in any case.

3. Getting information of other user
In this when user transfer from one location to another location at that time new location does not having rights to access data of that user at that time getting grant for accessing data of that user. When user's data is accessible from new location then it can -not access from old location.

3.3 System Functional Features

The cloud server provides the six main functions to the user.

1. Fine-Grained Access
In our propose system instead of showing complete data, the fetching of necessary data is allowed. Due to this system provides a quick response time.

2. Scalability
We are performing hierarchical structure so even if lower authority is absent for particular days at that time higher authority handle all work of lower authority so work of company will not be stopped.

3. Flexibility
When an employee gets transferred, his data could be accessible to the branch where he will be transferred only not to the older branch. So data will be transferred on request of CEO safely. Hence data can be transferred easily between branches.

4. Encryption
Encryption is a process in which data is hidden in a way that is accessible to the authorized user only. In this system we are providing encryption (converting into unreadable) so that data is not accessible by any illegal user like a hacker.

5. Decryption
Decryption is a process in which encrypted data i.e unreadable format is converted into readable format.

6. Key Generation and Verification
Key is generated based on the attributes filled by the user in registration form. In attribute key verification, when a key is used for login, it is first checked with the key stored in the database. If a match is found then user is allowed for further process else the user is rejected for further process.

4. ALGORITHM AND MATHEMATICAL MODEL

4.1 Algorithm

4.1.1RSA (Rivest Shamir Adleman)

Key generation

RSA involves a public key and a private key. The public key can be known by everyone and is used for encrypting

messages. Messages encrypted with the public key can only be decrypted in a reasonable amount of time using the private key.

Encryption

User1 transmits her public key *(n, e)* to User2 and keeps the private key secret. User1 then wishes to send message *M* to User2.He first turns *M* into an integer *m*, such that $0 \leq m < n$ by using an agreed-upon reversible protocol known as a scheme. He then computes the cipher text *c* corresponding to

$$c \equiv m^e \pmod{n}$$

This can be done quickly using the method of exponentiation by squaring. User1 then transmits *c* to User2.Note that at least nine values of *m* will yield a cipher text *c* equal to *m*, but this is very unlikely to occur in practice.

Decryption

User can recover m from c by using her private key exponent.

$$m \equiv c^d \pmod{n}$$

Given *m*, user can recover the original message *M* by reversing the padding scheme.

4.1.2 Advanced Encryption Standard Algorithm

The AES algorithm is also used for improving the searching and access mechanism.

4.2 Mathematical Model

We are using **NP-Complete** because it gives output within fix interval of time.

Set Theory Analysis

A] Identify the Employees

E= {e1, e2, e3….}

Where ‘E’ is main set of Employees like e1, e2, e3…

B] Identify the Attribute

AT= {at1, at2, at3….}

Where ‘AT’ is main set of registered Attribute like at1, at2, at3…

C] Identify the employee requested For another Attribute

RAA= {raa1, raa2, raa3}

Where ‘RAA’ is main set of Request for another Attribute raa1, raa2, raa3

D] Identify the employee requested for another employee Information

REI= {rei1, rei2, rei3}

Where ‘REI’ is main set of Request for another Attribute rei1, rei2, rei3

E] Identify Attribute Key of New employee

AK= {ak1, ak2, ak3….}

Where ‘AK’ is main set of attribute key of users ak1, ak2, ak3…

F] Identify the processes as P.

P= {Set of processes}
P = {P1, P2, P3,P4……}
P1 = {e1, e2, e3}
Where
{e1= upload data on server}

{e2= make the entry in database using different encryption algorithm}

{e3= get new attribute after request}

{e4= get new employee information when employee get transfer.}

G] Identify failure cases as FL

Failure occurs when –
FL= {F1, F2, F3…}
a) F1= = {f| ’f’ if error in uploading due to interrupted Internet connection}
H] Identify success case SS:-
Success is defined as-
SS= {S1, S2, S3, S4}

a) S1= {s|'s' if fast and no interrupted Internet connection}

b) S2= {s|'s' if data is added into database}

c) S2= {s|'s' if data is retrieve from database}

I] Initial conditions as I_0

a) User has good internet connection

b) Admin has good internet connection

H is universal set i.e cloud.

H= {E, B, U, R}

E=employee set

B=attribute set

U=user set

R=registered

A] Identify the Employees

E= {e1, e2, e3….}

Where 'E' is main set of Employees like e1, e2, e3…

B] Identify the Attribute

B= {at1, at2, at3….}

Where 'B' is main set of registered Attribute like at1, at2, at3…

C] Identify the employee requested For another Attribute

A= {raa1, raa2, raa3}

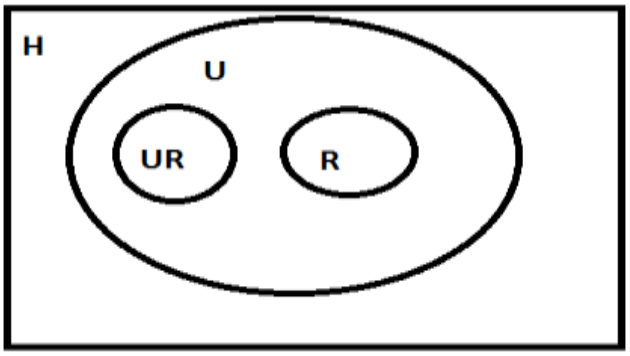
Where 'A' is main set of Request for another Attribute raa1, raa2, raa3

INITIAL STATE:

U={R, UR}

R=registered user

UR=unregistered user



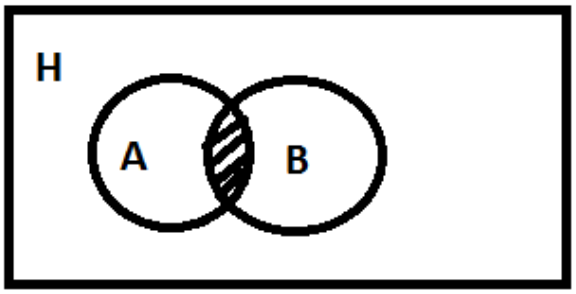
INTERMEDIATE STATE:

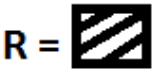
Request for new attribute

A=request for new attribute

B=contain all the attribute

R=provide requested attribute





S1=A ∩ B

Hierarchy

H = {H1, H2, H3, H4}

Where,

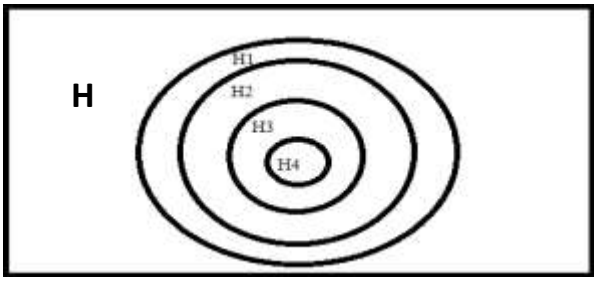
H is cloud

H1 is CEO.

H2 is general manager.

H3 is the list of managers.

H4 is the list of employees.



FLEXIBILITY:

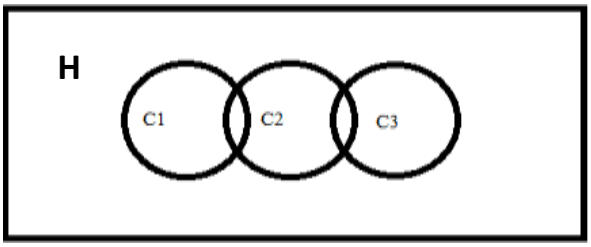
H= {C1, C2, C3}

Where,

C1 is the old branch of the company where employee worked before transfer.

C2 is the employee being transferred.

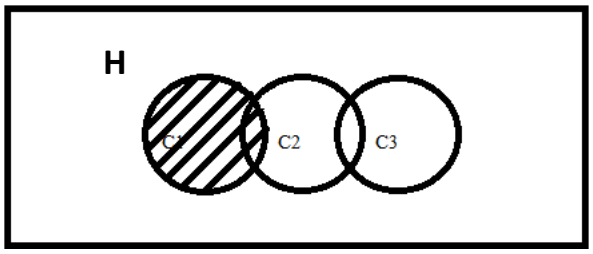
C3 is the new branch where employee got transferred to.



H= {C1, C2, C3}

Where,

S2 is employee data should be accessed to new branch only not old branch.



S2= (C1-C2) U C3

Scalability

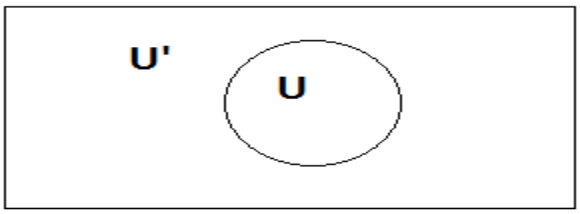
H = {H1, H2, H3, H4}

U= {H1, H2}

U'= {H3, H4}

U=present user

U'=absent user

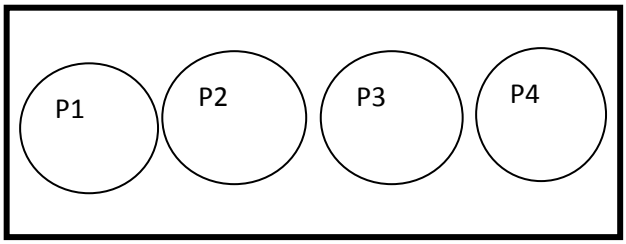


S3

FINAL STATE:

Identify the processes as P.

P= {Set of processes}
P = {P1, P2, P3, P4……}
Where
P1 = {S1, S2, S3}

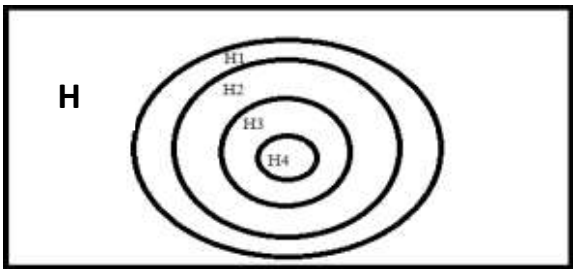


Where,

{S1= get new attribute after request}

{S2= get new employee information when employee get transfer.}

{S3= get access of lower authority}



FLEXIBILITY:

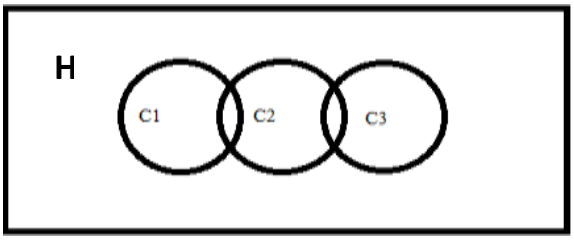
H= {C1, C2, C3}

Where,

C1 is the old branch of the company where employee worked before transfer.

C2 is the employee being transferred.

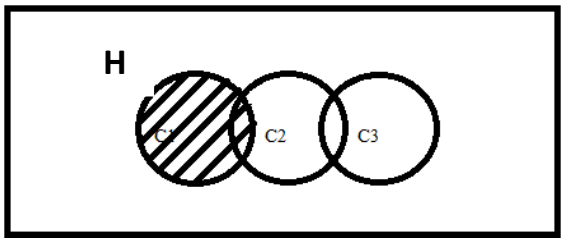
C3 is the new branch where employee got transferred to.



H= {C1, C2, C3}

Where,

S2 is employee data should be accessed to new branch only not old branch.



S2= (C1-C2) U C3

Scalability

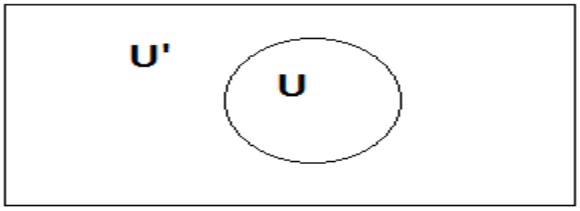
H = {H1, H2, H3, H4}

U= {H1, H2}

U'= {H3, H4}

U=present user

U'=absent user



S3

FINAL STATE:

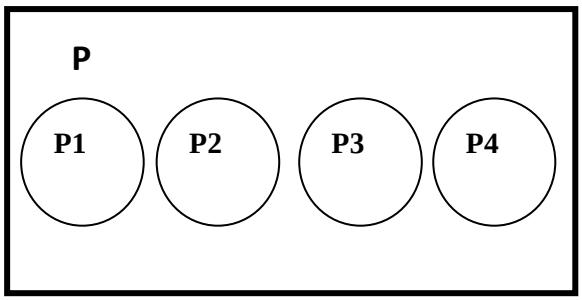
Identify the processes as P.

P= {Set of processes}

P = {P1, P2, P3, P4……}

Where

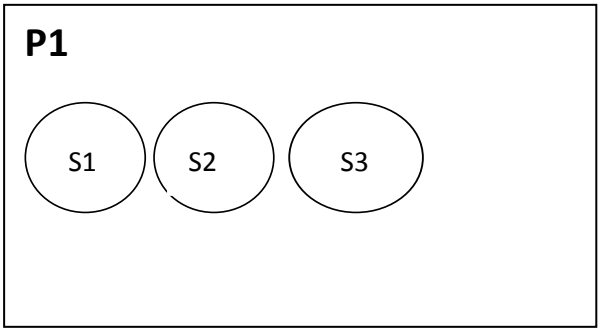
P1 = {S1, S2, S3}



{S1= get new attribute after request}

{S2= get new employee information when employee get transfer.}

{S3= get access of lower authority}



5. CONCLUSION

Thus, our system efficiently provides a fine grained access control with flexibility and scalability with a hierarchical structure in our HASBE system**.** Our paper will be providing security to the users from outsiders or intruders by implementing session hijacking and session fixation security in our system with sql injection attack prevention. The core is for sure, a cloud-base thus giving us a choice of multi-user access including security from intruder attacks. Hence we benefit the users with attack handling and many advantages over the existing systems.

6. REFERENCES

[1] Vipul et al.“ (Abe)Attribute based encryption”.

[2] Rakesh et al. “Attribute-Sets: A Practically Motivated Enhancement to Attribute-Based Encryption”, University of Illinois at Urbana-Champaign, July 27, 2009

[3] Pankaj et al. “Cloud Computing Security Issues in Infrastructure as a Service”,2012.

[4] John et al. “(cp-abe) Ciphertext-Policy Attribute-Based Encryption”John et al..

[5] Suhair et al. “Designing a Secure Cloud-Based EHR System using Ciphertext-Policy Attribute-Based Encryption”, 2011

[6] Ayad et al. ”Enabling Data Dynamic and Indirect Mutual Trust for Cloud Computing Storage System”, 2012.

[7]Chandana et al. “Gasbe: A Graded Attribute-Based Solution For Access Control In Cloud Computing”, 2011.

[8] Guojun et al. “Hierarchical attribute-based encryption and scalable user revocation for sharing data in cloud servers”, 2011.

[9] Qin et al. “Hierarchical Attribute-Based Encryption for Fine-Grained Access Control in Cloud Storage Services”.

[10] Patrick et al. “Methods and Limitations of Security Policy Reconciliation”.

[11]http://searchwindowsserver.techtarget.com/definition/IIS.

[12]http://en.wikipedia.org/wiki/Microsoft_Visual_Studio

[13] http://en.wikipedia.org/wiki/.NET_Framework.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 809 - 811, 2014, ISSN:- 2319–8656

CMS Website Security Threat Protection Oriented Analyzer System

Pritesh Taral
Department of Computer Engineering
Sinhagad Academy of Engineering,
Kondhwa (University of Pune)
Pune, Maharashtra, India

Balasaheb Gite
Department of Computer Engineering
Sinhagad Academy of Engineering
Kondhwa (University of Pune)
Pune, Maharashtra, India

Abstract - Website security is a critical issue that needs to be considered in the web, in order to run your online business healthy and smoothly. It is very difficult situation when security of website is compromised when a brute force or other kind of attacker attacks on your web creation. It not only consume all your resources but create heavy log dumps on the server which causes your website stop working.

Recent studies have suggested some backup and recovery modules that should be installed into your website which can take timely backups of your website to 3rd party servers which are not under the scope of attacker. The Study also suggested different type of recovery methods such as incremental backups, decremental backups, differential backups and remote backup.

Moreover these studies also suggested that Rsync is used to reduce the transferred data efficiently. The experimental results show that the remote backup and recovery system can work fast and it can meet the requirements of website protection. The automatic backup and recovery system for Web site not only plays an important role in the web defence system but also is the last line for disaster recovery.

This paper suggests different kind of approaches that can be incorporated in the WordPress CMS to make it healthy, secure and prepared web attacks. The paper suggests various possibilities of the attacks that can be made on CMS and some of the possible solutions as well as preventive mechanisms.
Some of the proposed security measures –

1. Secret login screen
2. Blocking bad boats
3. Changing db. prefixes
4. Protecting configuration files
5. 2 factor security
6. Flight mode in Web Servers
7. Protecting htaccess file itself
8. Detecting vulnerabilities
9. Unauthorized access made to the system checker

However, this is to be done by balancing the trade-off between website security and backup recovery modules of a website, as measures taken to secure web page should not affect the user's experience and recovery modules.

Keywords –WodrPress,Rsync.Web Security

1. INTRODUCTION

As WWW is becoming more and more complex lot of challenges has related to security of the webpage are arising. Website security is the most important part of the post development phase of the web creation. Web publisher needs to make check-ups of the websites and audit of the website to avoid the unexpected surprises. Website should be ready to withstand any attack made on it. Moreover, the website should not affect the user's experience and revenue by compromising the security of website.

It becomes a difficult situation when security of a website is compromised when any brute force attacker attacks on your creation. Attacker tries different permutations of password and username and it also consumes all your resources and create heavy log dumps on the server which causes your website stop working.

Sometimes attacker might get access to your website by injecting the code into website through open areas of the webpages such as comment box or any text field which is processed at the server side through php or any server side scripting language.

During holidays you don't have access to the administrator panel then you can put your website admin

panel into sleep mode so that no one can attack your login page.

Some of the proposed security measures are as follows–

1. Security for user accounts
2. Security for login module
3. Security while registering user
4. Security related to database module
5. htaccess and configuration file backup and restore
6. Functionality to blacklist and whitelist
7. Firewall protection and prevention of brute force login attack
8. whois lookup and security checker
9. Security for comment spam
10. Disabling access to source code and selection of text on UI

Backup can be taken using different approaches such as incremental backup, selective backup, complete backup and user can also recover from the hacking attack by using the restore mechanism which will restore system to previous working state. Backup can be complete database backup.

This paper basically deals with mechanisms mentioned above to secure website from bad boats and hackers and make your server healthy by removing possible security threats. The Paper also pesents different backup and restore mechanisms.

2. RELATED WORK

There has been extensive efforts made to understand web security by considering network traffic, encryption techniques etc. But very few efforts have been taken to understand the security needs of CMS and the techniques to deal with them.

Some of the important work related with this study is as follows:

A web site protection oriented remote backup and recovery method :

He Qian, Guo Yafeng, Wang Yong, in his thesis describes that how we can take incremental abd decremental backups of the website which will be used to recover site during disaster. [1].

Website Regional Rapid Security Detection Method:

Yong Fang ; Liang Liu, suggested that distributed design, website regional rapid security detection method can conduct security detection for the whole region by adding detection module dynamically as needed and record results of detection. [2]

Research and solution of existing security problems in current internet website system :

Guiyang ; Xiaoyao Xie analyses the network system faced by common threats and attack methods and means for the typical, sum-up a website security system by the need to address the problem, solve these problems formulate the corresponding protection measures.

4. SECURITY MEASURES

Security for user account

Sometimes CMS might have user account with default user name ‘admin’ which is easier for attacker to predict and attack or query to your CMS. It is considered as bad security practice as it makes the task of attacker 50% easier because attacker already knows one of your credentials required to login. Besides this a Password strength tool can be used to allow you to create very strong passwords.

Security for login module

It is to protect the CMS against brute force login attacks with the login lockdown feature so that users with certain IP range can be locked out of the system for a predetermined amount of time based on the configuration setting. It also force logout of all users after a configured period of time

1. **Security while registering user**
   Enable manual approve feature for registered accounts can minimize spam and bogus registrations. Captcha can also help us to prove valid user.

2. **Security related to database module**
   Table prefixes can be modified to other prefixes to make security level higher for the attacker. Attacker cannot predict the table prefix much easily

3. **htaccess and configuration file backup and restore**
   Configurations files which are useful for running website should be protected from attacks. It is main file which provides security to other CMS modules.

4. **Functionality to blacklist and whitelist**
   It is used to blacklist and whitelist IP addresses of the web surfers. It is recommended to identify the search engine boat and spam boats

4. ANALYZE AND SUGGEST TOOL

Analyze and suggest tool is used to scan the CMS website for checking out possible threat inside the system. It then analyze the website and generate the security

reports and suggest out some possible solutions and also provides option to incorporate them into the current CMS system

Figure 1: General Architecture of the System

5. CONCLUSION

CMS security is quite different from the traditional notions of website security. CMS has a predefined structure and it is used by millions of peoples to create websites. This fact makes the attacker's task easy, as he already knows the predefined structure of CMS. Our concept would modify the traditional CMS structure into a new customized CMS so that the structure of the system would not remain as a default. Thus it becomes difficult for an attacker to predict the DB and configuration structure of the CMS which would eventually boost the security level in CMS up.

.

6. REFERENCES

[1] He Qian, Guo Yafeng, Wang Yong, Qiang Baohua1 "A web site protection oriented remote backup and recovery method" INSPEC Accession Number : 14022497 2014 IEEE

[2] Yong Fang ; Liang Liu, "Website Regional Rapid Security Detection Method" 978-1-4799-0587-4 20 13 IEEE

[3] Gaoqi Wei, "Research and solution of existing security problems in current internet website system", 978-1-4244-2584-6 20 08 IEEE

[4] Wenping Liu ; Xiaoying Wang ; Li Jin, "Design and implementation of a website security monitoring system from users' perspective", 978-0-7695-4817-3/12 2012 IEEE

Pritesh A.Taral received the B.E. degree in Computer Engineering from S.A.O.E Pune, INDIA in 2011 and perusing M.E. degree in Computer Engineering from S.A.O.E , Pune

Prof. Balasaheb B.Gite is working as Head of the department of Computer engineering at SAOE Pune India. He received the B.E. degree in Computer Engineering from P.R.E.C Loni INDIA and M.E. degree from W.C.E, Pune.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 812 - 814, 2014, ISSN:- 2319–8656

Sentence Validation by Statistical Language Modeling and Semantic Relations

Lakshay Arya
Guru Gobind Singh Indraprastha University
Maharaja Surajmal Institute Of Technology
New Delhi, India

Abstract : This paper deals with Sentence Validation - a sub-field of Natural Language Processing. It finds various applications in different areas as it deals with understanding the natural language (English in most cases) and manipulating it. So the effort is on understanding and extracting important information delivered to the computer and make possible efficient human computer interaction. Sentence Validation is approached in two ways - by Statistical approach and Semantic approach. In both approaches database is trained with the help of sample sentences of Brown corpus of NLTK. The statistical approach uses trigram technique based on N-gram Markov Model and modified Kneser-Ney Smoothing to handle zero probabilities. As another testing on statistical basis, tagging and chunking of the sentences having named entities is carried out using pre-defined grammar rules and semantic tree parsing, and chunked off sentences are fed into another database, upon which testing is carried out. Finally, semantic analysis is carried out by extracting entity relation pairs which are then tested. After the results of all three approaches is compiled, graphs are plotted and variations are studied. Hence, a comparison of three different models is calculated and formulated. Graphs pertaining to the probabilities of the three approaches are plotted, which clearly demarcate them and throw light on the findings of the project.

Keywords: language modeling, smoothing, chunking, statistical, semantic

1. INTRODUCTION

NLP is a field of Computer Science and linguistics concerned with interactions between computers and human languages. NLP is referred to as AI-complete problem. Research into modern statistical NLP algorithms require understanding of various disparate fields like linguistics, computer science, statistics, linear algebra and optimization theory.

To understand NLP, we have to keep in mind that we have several types of languages today : Natural Languages such as English or Hindi, Descriptive Languages such as DNA, Chemical formulas etc, and artificial languages such as Java, Python etc. We define Natural Language as a set of all possible texts, wherein each text is composed of sequence of words from respective vocabulary. In essence, a vocabulary consists of a set of possible words allowed in that language. NLP works on several layers of language: Phonology, Morphology, Lexical, Syntactic, Semantic, Discourse, Pragmatic etc. Sentence Validation finds its applications in almost all fields of NLP - Information Retrieval, Information Extraction, Question-Answering, Visualization, Data Mining, Text Summarization, Text Categorization, Machine and Language Translation, Dialogue And Speech based Systems and many other one can think of.

Statistical analysis of data is the most popular method for applications aiming at validating sentences. N-gram techniques make use of Markov Model. For convenience, we restrict our study till trigrams which are preceded by bigrams. Results of this approach are compared with results of Chunked-Off Markov Model. Extending our study and moving towards Semantic Analysis - we find out the Entity-Relation pairs from the Chunked off bigrams and trigrams. Finally, we aim to calculate the results for comparison of the three above models.

2. SENTENCE VALIDATION

Sentence validation is the process in which computer tries to calculate the validity of sentence and gives the cumulative probability. Validation refers to correctness of sentence, in dimensions such as statistical and semantic. A good validation program can verify whether sentence is correct at all levels.

Python language and its NLTK [5] suite of libraries is most suited for NLP problems. They are used as a tool for most of NLP related research areas - empirical linguistics, cognitive science, artificial intelligence, information retrieval and machine learning. NLTK provides easily-guessable method names for word tokenizing, sentence tokenizing, POS tagging, chunking, bigram and trigram generation, frequency distribution, and many more. Oracle connectivity with Python is used to store the bigrams, trigrams and entity-relation pairs required to test the three different models and finally to compare their results.

First model is the purely statistical Markov Model, i.e. bigrams and trigrams are generated from the sample files of Brown corpus of NLTK and then fed into the database. Testing yields some results and raises some disadvantages which will be discussed later. Second model is Chunked-Off Markov Model - an extension of the first model in the way that it makes use of tagging and chunking operations wherein all the proper nouns are categorized as PERSON, PLACE, ORGANIZATION, FACILITY, etc. This replacing solves some issues which purely statistical model could not deal with. Moving from statistical to semantic approach, we now aim to validate a sentence on semantic basis too, i.e. whether

the sentence has some meaning and makes sense or not. For example, 'PERSON eats' is a valid sentence whereas 'PLACE eats' is an invalid one. So the latter input sentence must result in a low probability for correctness compared to the former. In order to show this demarcation between sentences, we extract the entity relation pairs from sample sentences using named entity recognition and chunking and store them in the ER database. Whenever a sentence comes up for testing, we

extract the E-R pairs in this sentence and match them from database entries to calculate probability for semantic validity.

The same corpus data and test data for the above three approaches are taken for comparison purposes. Graphs pertaining to the results are plotted and major differences and improvements are seen which are later illustrated and analyzed.

3. HOW DOES IT WORK ?

The first two statistical approaches use the N-gram technique and Markov Model[2] building. In the pure statistical Markov N-gram Model, corpus data is fed into the database in the form of bigrams and trigrams with their respective frequencies(i.e. how many times they occur in the whole data set of sample sentences). When an input sentence is to be validated, it is tokenized into bigrams and trigrams which are then matched with database values and a cumulative probability after application of Smoothing-off technique of Kneser-Ney Smoothing which handles new words and zero count events having zero probability which may cause system crash, is calculated.

Chunked-Off Markov Model makes use of our own defined replace function implemented through pos_tag and ne_chunk functionality of NLTK. Every sentence is first tagged according to Part-Of-Speech using pos_tag. Whenever a 'NN', 'NNP' or in general 'NN*' chunk is encountered, it is passed to ne_chunk which replaces the named entity with its type and returns a modified sentence whose bigrams and trigrams are generated and fed into the database. The testing procedure of this approach follows above methodology and modifies the sentence entered by the user in the same way, calculates the probabilities of the bigrams and trigrams by matching them with database entries and finally smoothes off to yield final results.

Above two approaches are statistical in nature, but we need to validate sentences on semantic and syntactic basis as well, i.e. whether sentences actually make sense or not. For bringing this into picture, we extract all entities(again NN* chunks) and relations(VB* chunks). We define our own set of grammar rules as context free grammar to generate parse tree from which E-R pairs are extracted.

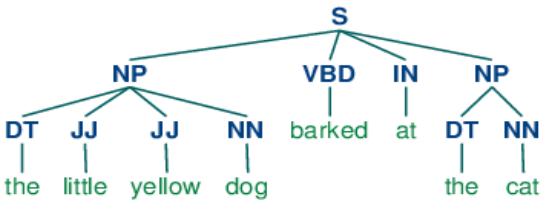


Figure. 1 Parse Tree Generated by CFG

4. COMPLETE STRUCTURE

We have trained the database with 85% corpus and testing with the rest of 15% corpus we have. This has two advantages - firstly we shall use the same ratio in all other approaches so that we can compare them easily. Secondly it provides a threshold value for probability which will help us to distinguish between correct and incorrect test sentences depicting regions above and below threshold respectively. Graphs are plotted between probability(exponential, in order of 10) and length of the sentence(number of words).

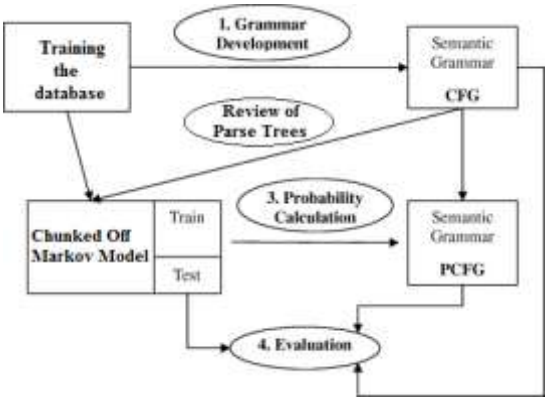


Figure. 2 Complete flowchart of Sentence Validation process

4.1 N-Gram Markov Model

The first module is Pure Markov Model[1]. In the pure statistical Markov N-gram Model, corpus data is fed into the database in the form of bigrams and trigrams with their respective frequencies(i.e. how many times they occur in the whole data set of sample sentences). When an input sentence is to be validated, it is tokenized into bigrams and trigrams which are then matched with database values and a cumulative probability after application of Smoothing-off technique of Kneser-Ney Smoothing is calculated. The main disadvantage of this pure statistics-based model is that it is not able to deal with Proper Nouns and Named Entities. Whenever a new proper noun is encountered with the same relation, it will result in lower probability even though the sentence might be valid. This shortcoming of Markov Model is overcome by next module - Chunked Off Markov Model. Markov Modeling is the most common method to perform statistical analysis on any type of data but it cannot be the sole model for testing of NLP applications.

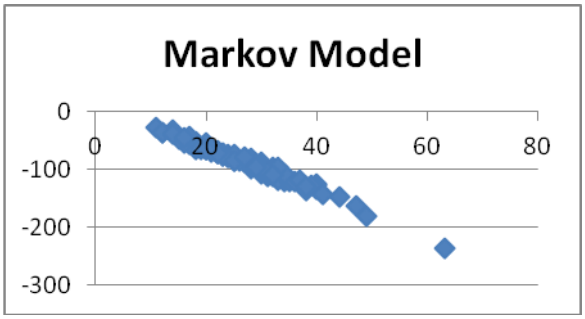


Figure. 3 Testing results for Pure Statistical Markov Model

4.2 Chunked-Off Markov Model

The second module is Chunked-Off Markov Model[3] - training the database with corpus sentences in which all the nouns and named entities are replaced with their respective type. This is implemented using the tagging and chunking operations of NLTK. This solves the problem of Pure Statistical model that it is not able to deal with proper nouns. For example, a corpus sentence has the trigram 'John eats pie'. If a test sentence occurs like 'Mary eats pie', it will result in a very low trigram probability. But if the trigram 'John eats pie' is modified to 'PERSON eats pie', it will result in a better comparison.

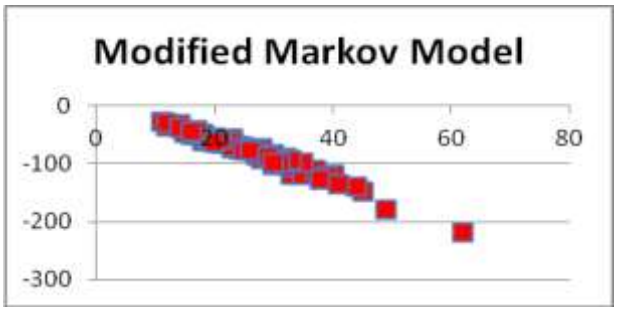


Figure. 4 Testing results for Chunked-Off Markov Model

4.3 Entity-Relation Model

The third module is E-R model[4]. Extraction of the entities(all NN* chunks) and relations(all VB* chunks). We define a set of grammar rules as context free grammar to generate parse tree from which E-R pairs are extracted and entered into the database. For convenience, we have taken the first main entity and the first main relation because compound entities are difficult to deal with.

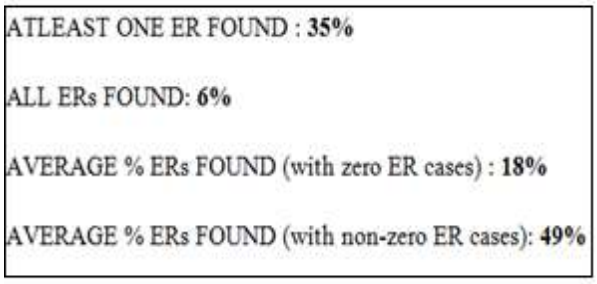


Figure. 5 Testing results for E-R Model

4.4 Comparison of the three models

The fourth module is comparison. As expected, the modified Markov Chunked-Off model performs. We can also see that there are no sharp dips in the modified model which are present in pure statistical model due to a sharp decrease in the probability of trigrams and bigrams. The modified model is consistent due to its ability to deal with Proper Nouns and Named Entities.

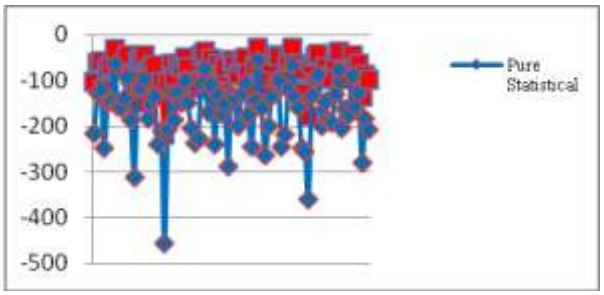


Figure. 6 Comparison of the two models

5. WHY SENTENCE VALIDATION ?

Sentence Validation finds its use in the following fields :-

1. Information systems
2. Question-answering systems
3. Query-based information extraction systems
4. Text summarization applications
5. Language and machine translation applications
6. Speech and Dialogue based systems

For example, Wolfram-Alpha, Google Search Engine, Text Compactor, SDL Trados, Siri, S-Voice, etc all integrate sentence validation as an important module.

6. CHALLENGES AND FUTURE OF SENTENCE VALIDATION

As mentioned earlier NLP is still in the earliest stage of adoption. Research work in this field is still emerging. It is a difficult task to train a computer and make it understand the complex and ambiguous nature of natural languages. The statistical approach is a well proven approach for statistical calculations. But the data obtained from ER approach is inconclusive. We may have to improve our approach and scale the data to make ER model work. ER Model offers very substantial advantages over Statistical Model, that makes this approach worth looking into. Even if it cannot reach the levels of Markov Model, ER Model could be a powerful tool in complementing Markov Model as well as for variety of other NLP Applications.

We see Sentence Validation as the single best method available to process any Natural Language application. All languages have own set of rules which are not only difficult to feed in a computer, but are also ambiguous in nature and complex to comprehend and generalize. Thus, different approaches have to be studied, analyzed and integrated for accurate results.

Our three approaches validate a sentence in an over-all manner, both statistically and semantically, making this system an efficient one. Also, the graphs show clearly that chunking of the training data will yield in better testing of data. The testing will become even more accurate if database is expanded with more sentences.

7. REFERENCES

[1] Chen, Stanley F. and Joshua Goodman. 1998 An empirical study of smoothing techniques for language modeling Computer Speech & Language 13.4 : 359-393.

[2] Goodman. 2001 A bit of progress in Language Modeling

[3] Rosenfield, Roni. 2000 Two decades of statistical language modeling : Where do we go from here?

[4] Nguyen Bach and Sameer Badaskar. 2005 A Review of Relation Extraction

[5] NLTK Documentation. 2014 Retrieved from http://www.nltk.org/book

International Journal of Computer Applications Technology and Research
Volume-3,Issue-12, December 2014

Implementation of Adaptive Digital Beamforming using Cordic

AZRA JEELANI
Associate Professor,
M S Engineering College,
Bangalore, Karnataka, India
azrajeelani@gmail.com

Dr. VEENA.M.B
Associate Professor,
B M S College of Engineering,
Bangalore, Karnataka, India.
veenamb.ece@bmsce.com

Dr. CYRIL PRASANNA RAJ
Dean R & D,
M S Engineering College,
Bangalore, Karnataka, India
cyril@msec.ac.in

Abstract: Sonar imaging is one of the simplest technique for detection of under water drowned bodies. There is a need for design of conventional beamforming which are robust and simple. Adaptive beamformer is used to improve the quality of the sonar image. As a result we get an image containing more useful and correct information. The CORDIC computing technique a highly efficient method to compute elementary functions like sine, cosine, translate, rotate values using CORDIC algorithm. The system simulation was carried out using ModelSim and Xilinx ISE Design Suite 9.2i.. Matlab code is used to implement sin and cos using cordic angles and amplitude response of beamformed data by optimized method in order to enlarge the validity region of beamforming. Synthesis results of cordic shows the reduced memory requirement and less power consumption.

Keywords: Beamforming , cordic, sonar imaging, validity region

1. INTRODUCTION

Beamforming is a type of signal processing technique used in sensor arrays for directional signal transmission or reception. Here the elements are combined in such a way that signals at particular angles experience constructive interference while others experience destructive interference. 3D sonar imaging has been one of the main innovations in underwater applications over the years[1]. There are two critical issues in the development of high resolution 3D sonar systems are 1) the cost of hardware, which is associated with the huge number of sensors that compose the planar array and 2) the computational burden in processing the signals. Palmese and Trucco also propose an algorithm to perform chirp zeta transform beam forming on the wideband signals collected by an evenly spaced planar array and generated by a scene placed in both the far field and the near field [4],[6]. Works are done in [8]-[10] have proposed to use the Coordinated Rotation DIgital Computer(CORDIC) in implementing frequency domain beamforming on field Programmable Gate Arrays-the CORDIC algorithm in an iterative arithmetic algorithm given by Volder[11] and Walther[12].This paper describes a data path using CORDIC for the algorithm.

The digital signal processing has long been dominated by microprocessors with enhancements such as single cycle multiply-accumulate instructions and special addressing modes. While these processors are low cost and offer extreme flexibility, they are not fast enough for truly demanding DSP tasks. The advent of high speeds of dedicated hardware solutions which has the costs that are competitive with the traditional software approach. Unfortunately, algorithms optimized for these microprocessor based systems do not always map well into hardware. While hardware-efficient solutions often exist, the dominance of the software systems has kept those solutions out of the spotlight. Among these hardware-efficient algorithms is a class of iterative solutions for trigonometric and other functions that use only shifts and adds to perform. The trigonometric functions are based on vector rotations, while other functions like square root are implemented using an incremental expression of the desired function. The trigonometric algorithm is called CORDIC, an acronym for COordinate Rotation DIgital Computer. The incremental functions are performed with a very simple extension to the hardware architecture, and while not CORDIC in the strict sense, are often included because of the close similarity. The CORDIC algorithms generally produce one additional bit of accuracy for each iteration. The trigonometric CORDIC algorithms were originally developed as a digital solution for real-time navigation problems. The original work is credited to Jack Volder [4,9]. Extensions to the CORDIC theory based on work by John Walther[1] and others provide solutions to a broader class of functions. This paper attempts to survey the existing CORDIC and CORDIC-like algorithms and then towards implementation in Field Programmable Gate Arrays (FPGAs).

A approximation of used in near field beamforming presented in[13],[14] by enlarging the validity region. Beamforming can be used at both the transmitting and receiving ends in order to achieve spatial selectivity. To change the directionality of the array when transmitting, a beamformer controls the phase and relative amplitude of the signal at each transmitter, in order to create a pattern of constructive and destructive interference in the wave front. When receiving, information from different sensors is combined in a way where the expected pattern of radiation is preferentially observed. Conventional beamformers use a fixed set of weightings and time-delays (or phasings) to combine the signals from the sensors in the array, primarily using only information about the location of the sensors in space and the wave directions of interest. In contrast, adaptive beamforming techniques generally combine this information with properties of the signals actually received by the array, typically to improve rejection of unwanted signals from other directions. This process may be carried out in either the time or the frequency domain. Hardware implementation of bio-inspired algorithm for motion detection takes less processing time. Integration of motion detection model and improves the performance of autonomous visual navigation. For resolving the navigation problems two existing approach optical flow or non bio-inspired and bio- inspired processing time is needed to reduce. For minimizing the size of system algorithm should be implemented on ASIC and functionality should be verified on

FPGA before taking to ASIC.

2. BACKGROUND THEORY:

2.1.Beamforming:

Beamforming is a type of signal processing technique used in sensor arrays for directional signal transmission or reception. Here the elements are combined in such a way that signals at particular angles experience constructive interference while others experience destructive interference[1]. Beamformers are classified as either data independent or statistically optimum, depending on how the weights are chosen. The weights in a data independent beamformer do not depend on the array data and are chosen to present a specified response for all signal and interference scenarios. The weights in a statistically optimum beamformer are chosen based on the statistics of the array data to optimize the array response. The statistics of the array data are not usually known and may change over time so adaptive algorithms are typically used to determine the weights. The adaptive algorithm is designed so the beamformer response converges to a statistically optimum solution [6].

The weights in a data independent beam former are designed so that the beamformer response approximates a desired response independent of the array data or data statistics. This design objective is same as that for a classical FIR filter design. The simple delay and sum beam former is an example of the data independent beamforming.

In statistically optimum beam former the weighs are chosen based on the statistics of the data received at the array. The goal is to optimize the beam former response so that the output signal contains minimal contributions due to the noise and signals arriving from directions other than the desired direction. The Frost beamformer is a statistically optimum beam former. Other statistically optimum beamformers are Multiple Side lobe Canceller and Maximization of the signal to noise ratio.

2.2.Sonar Imaging:

Sonar (an acronym for SOund Navigation and Ranging) is a technique that uses sound propagation (usually underwater, as in submarine navigation) to navigate, communicate with or detect objects on or under the surface of the water, such as other vessels. Two types of technology share the name "sonar": passive sonar is essentially listening for the sound made by vessels; active sonar is emitting pulses of sounds and listening for echoes. Sonar may be used as a means of acoustic location and of measurement of the echo characteristics of "targets" in the water. Acoustic location in air was used before the introduction of radar. Sonar may also be used in air for robot navigation, and SODAR (upward looking in-air sonar) is used for atmospheric investigations. The term sonar is also used for the equipment used to generate and receive the sound.

2.3.Active and Passive Sonar System

Active sonar or passive sonar, when receiving the acoustic signal reflected from the target, the information included in the signal cannot be directly collected and used without technical signal processing. To extract the efficient and useful information's from the mixed signal, some steps should be taken to transfer sonar data from raw acoustic data reception to detection output.

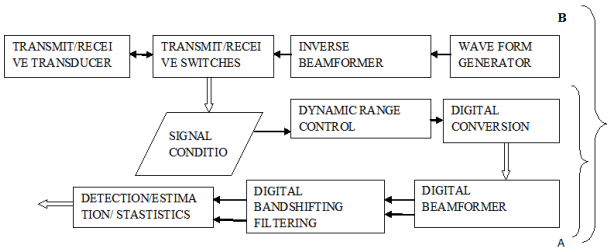


Fig.1 Passive and Acive Sonar System

$A = \pi r^2$ needed during the signal processing system as shown in Fig.1

2.4.Cordic Theory:

Coordinate Rotational Digital Computer (CORDIC) is a set of shift-add algorithm known for computing a wide range trigonometric functions, hyperbolic, linear and logarithmic functions also like multiplication division, data type conversion, square root. It is highly efficient, low complexity. The CORDIC algorithm has found in various applications such as pocket calculator, numerical co-processors to high performers Radar signal processing, supersonic bomber. Vector rotation can also be used for polar to rectangular and rectangular to polar conversions, for vector magnitude, and as a building block in certain transforms such as the DFT and DCT. The CORDIC algorithm provides an iterative method of performing vector rotations by arbitrary angles using only shifts and adds. The algorithm, credited to Volder[4], is derived from the general (Givens) rotation transform:

$$x' = x.\cos(\phi) - y.\sin(\phi) \quad \text{----------(1)}$$

$$y' = x.\cos(\phi) + y.\sin(\phi) \quad \text{----------(2)}$$

which rotates a vector in a Cartesian plane by the angle ϕ. These can be rearranged so that:

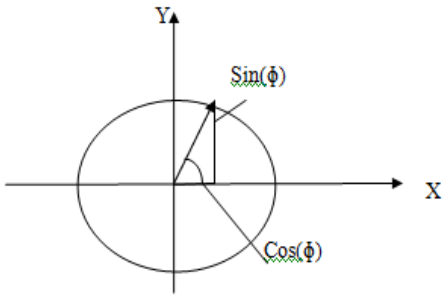


Fig 2. Rotation of sin and cos

$$x' = \cos(\phi).[x - y.\tan(\phi)] \quad \text{----------(3)}$$

$$y' = \cos(\phi).[x + y.\tan(\phi)] \quad \text{----------(4)}$$

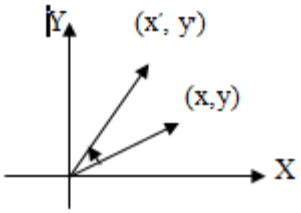


Fig 3. Input and output of rotation for rotation mode

So far, nothing is simplified. However, if the rotation angles are restricted so that $\tan(\phi)2^{-i}$, the multiplication by the tangent term is reduced to simple shift operation. Arbitrary angles of rotation are obtainable by performing a series of successively smaller elementary rotations. If the decision at each iteration, i, is which direction to rotate rather than whether or not to rotate, then the $\cos(\delta_i)$ term becomes a constant (because $\cos(\delta_i) = \cos(-\delta_i)$). The iterative rotation can now be expressed as:

$$X_{i+1} = K_i[\, x_i - d_i.y_i.2^{-i}\,] \quad \text{---------(5)}$$

$$Y_{i+1} = K_i[\, x_i + d_i.y_i.2^{-i}\,] \quad \text{--------(6)}$$

Where

$$K_i = \cos(\tan^{-1}2^{-i}) = 1/\mathrm{sqrt}(1+ 2^{-2i}) \quad \text{----(7)}$$

$$d_i = \pm 1 \quad \text{----------(8)}$$

Removing the scale constant from the iterative equations yields a shift-add algorithm or vector rotation. The product of the K_i's can be applied elsewhere in the system or treated as part of a system processing gain. That product approaches 0.6073 as the number of iterations goes to i infinity. Therefore, the rotation algorithm has a gain, A_n of approximately 1.647. The exact gain depends on the number of iterations, and obeys the relation

$$A_n = \prod \mathrm{sqrt}(1+ 2^{-2i}) \quad \text{----------(9)}$$

The angle of a composite rotation is uniquely defined by the sequence of the directions of the elementary rotations. That sequence can be represented by a decision vector. The set of all possible decision vectors is an angular measurement system based on binary arctangents. Conversions between this angular system and any other can be accomplished using look-up. A better conversion method uses an additional adder-subtractor that accumulates the elementary rotation angles at each iteration. The elementary angles can be expressed in any convenient angular unit. Those angular values are supplied by a small lookup table (one entry per iteration) or are hardwired, depending on the implementation. The angle accumulator adds a third difference equation to the algorithm.

$$Z_{i+1} = Z_i + d_i(\tan^{-1} 2^{-i}) \quad \text{----------(10)}$$

Obviously, in cases where the angle is useful in the arctangent base, this extra element is not needed. The CORDIC rotator is normally operated in one of two modes. The first, called rotation by Volder[4], rotates the input vector by a specified angle (given as an argument). The second mode, called vectoring, rotates the input vector to the *x* axis

2.5. Implementation in an FPGA

There are a number of ways to implement a CORDIC processor. The ideal architecture depends on the speed versus area tradeoffs in the intended application. First we will examine an iterative architecture that is a direct translation from the CORDIC equations. From there, we will look at a minimum hardware solution and a maximum performance solution.

2.6.Iterative CORDIC Processors

An iterative CORDIC architecture can be obtained simply by duplicating each of the three difference equations in hardware as shown in Figure 1. The decision function, d_i, is driven by the sign of the *y* or *z* register depending on whether it is operated in rotation or vectoring mode. In operation, the initial values are loaded via multiplexers into the x, y and z registers. Then on each of the next n clock cycles, the values from the registers are passed through the shifters and adder-subtractors and the results placed back in the registers. The shifters are modified on each iteration to cause the desired shift for the iteration. Likewise, the ROM address is incremented on each iteration so that the appropriate elementary angle value is presented to the z adder-subtractor. On the last iteration, the results are read directly from the adder-subtractors. Obviously, a simple state machine is required keep track of the current iteration, and to select the degree of shift and ROM address for each iteration. The design depicted in Figure 1 uses word-wide data paths (called bit-parallel design). The bit-parallel variable shift shifters do not map well to FPGA architectures because of the high fan-in required. If implemented, those shifters will typically require several layers of logic (i.e., the signal will need to pass through a number of FPGA cells). The result is a slow design that uses a large number of logic cells.

3. PROPOSED WORK:

Digital input pulse is passed to find the angle or detection of object under water. In Fig.3.Beamforming can be used at both the transmitting and receiving ends in order to achieve spatial selectivity, the data is transmitted to underwater sonar system, sonar is used to detect the underwater objects and finds the angle elevation. The beamformed data is transmitted; at the receiver end beam formation data is generated. The generated beam formation data will be having interference and noise error that will be reduced by using optimization technique. Optimized cordic beamforming will eliminate all the interference which generated at receiver end. Final optimization beamforming data is obtained

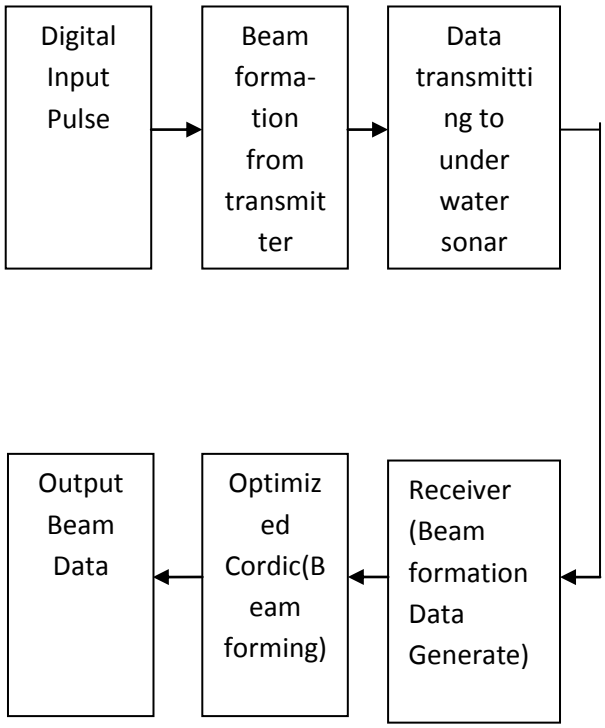


Fig 4.. Beamforming for underwater sonar

3.1.Program Flow Chart:

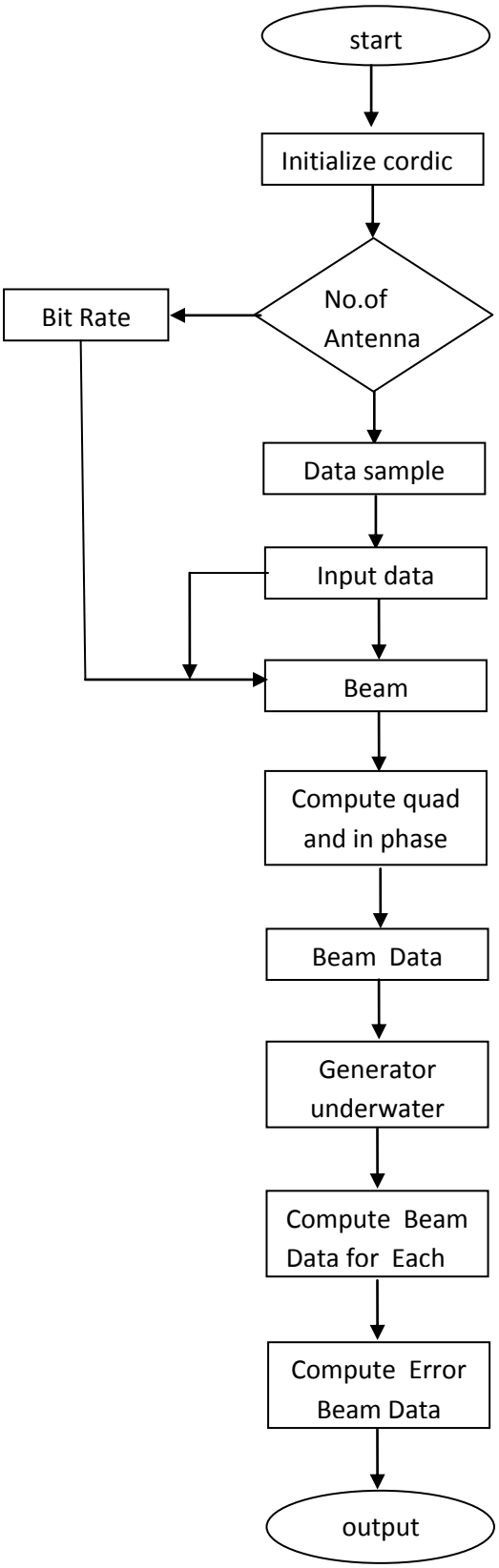


Fig.5 .Program Flow Chart

Fig 5. shows the flow chart in which initially the CORDIC values are sampled, the antennas are used to detect the angles and object of beam form underwater sonar with sampled bit rate. The detected angles are taken as input data where the beam data is formed, the obtained beam data are sampled according to mathematical calculations under CORDIC algorithm, the obtained beam data samples are computed as quad phase and In phase. The received beam data contains noise and interference which are reduced and eliminated using underwater noise model. The beam data is computed for each antenna and its angles, the error beam data is finally computed to obtain noiseless beam data. The obtained output is in the form of optimized beam form data.

3.2. Architecture:

The architecture is shown in Fig.6 in which input signals are given to memory. The memory is used to store the data of input signals. The signals are transmitted to detect the target or object in underwater beam form data. Once the target is detected and beam form data is generated. The received beam form data is up sampled and Down sampled. The adder is used to combine the images received and stored in memory. The generated beam form data signals from sonar are given to CORDIC algorithm

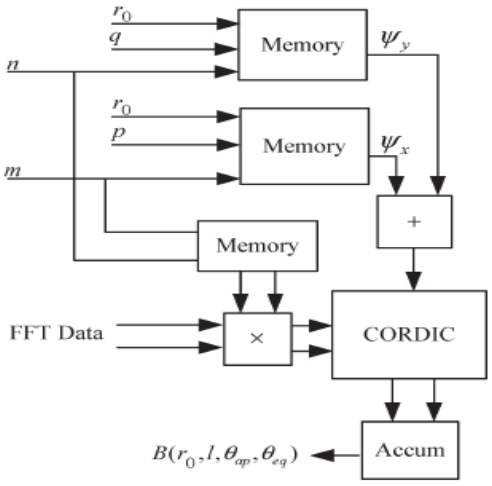


Fig.6. Data path algorithm

The received data is sampled according to CORDIC algorithm calculations. The Angle is measured using CORDIC. The sin and cos angles are generated and calculated using CORDIC algorithm. Both the IN Phase and Quad Phase is added and given to Cordic using adder. The CORDIC performs vector rotation and the vector data are some to produce the array beam (B). Angles which are detected are to be measured using CORDIC. The obtained Samples are stored in the register.

Rotation vector is given by equation (3 and (4)

To find iterations the following equations are used

$$X_{i+1}\ \alpha_i = x_i - d_i.y_i.2^{-i} \quad \text{----------(11)}$$

$$Y_{i+1}\ \alpha_i = x_i + d_i.y_i.2^{-I} \quad \text{----------(12)}$$

$$Z_{i+1}\alpha_i = x_i - d_i.\arctan(2^{-i}) \quad \text{-----} \quad \text{-------(13)}$$

To find magnitude and phase the following equations are used

$$X' = Z_n \sqrt{(X^2+Y^2)} \quad \text{----------(14)}$$

$$Y' = 0 \quad \text{----------(15)}$$

$$\Theta' = \text{atan}(x/y) \quad \text{----------(16)}$$

Advantages and Disadvantages of CORDIC

- Simple Shift-and-add Operation.(2 adders+2 shifters vs. 4 mul.+2 adder)
- It needs *n* iterations to obtain *n*-bit precision.
- Slow carry-propagate addition.
- Low throughput rate and area consuming shifting operations.

The m and n are the input coordinates, p and q are pre-computed values, r_o are the rotation value ,ψ_x and ψ_y is the phase shifter. By using L point DFT the sample data are calculated by

$$S_{m,n}(l) = \sum S_{m,n}(t) \exp(-j*2*\prod *t*l)/L \ ,$$

$$t=0 \text{ to } L-1 \quad \text{----------(17)}$$

The $S_{m,n}(l)$ are stored in memory with indexing parameter m and n .The phase shift parameter ψ_x and ψ_y is added to the phase term of the data $W_{m,n}.S_{m,n}(l)$ by CORDIC which perform as a vector rotation. The vector data is summed to produce the array beam of $B(r_o,\Theta_{ap},\Theta_{eq})$ where l is the frequency , Θ_{ap} and Θ_{ep} is the time delay.

4. RESULTS & DISCUSSIONS:

4.1 Results of Direct Method Using MATLAB

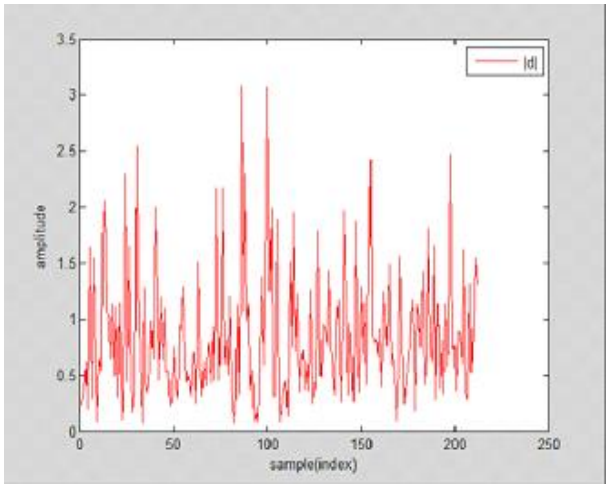


Fig.7 Amplitude of Transmitted Data

Fig.7 shows the variation based on the size of the data from the transmitter side. Fig.8 shows the phase wise changes from -10 degree to -40 degree based on the optimized algorithm.

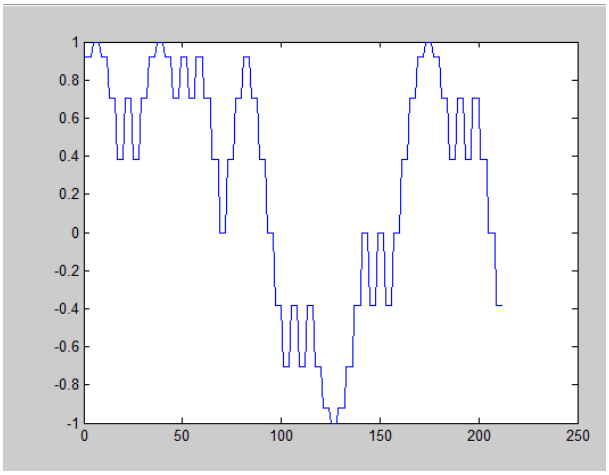


Fig.8. Input Data

In Fig.8 the Phase Graph of Input Data is transmitted.

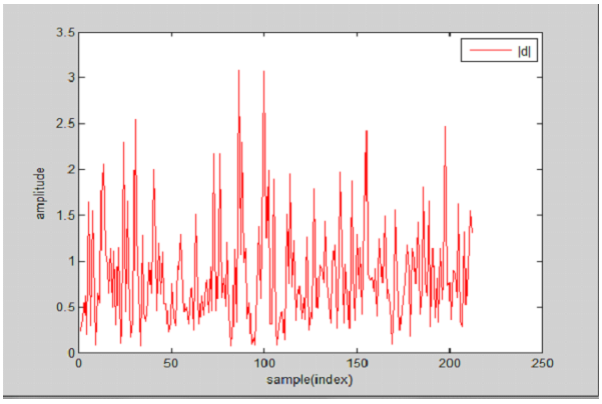


Fig.9 Amplitude of Transmitted data.

As shown in Fig 9. the amplitude variation based on the size of the data from the transmitter side and the amplitude variation generate from the beamforming data.

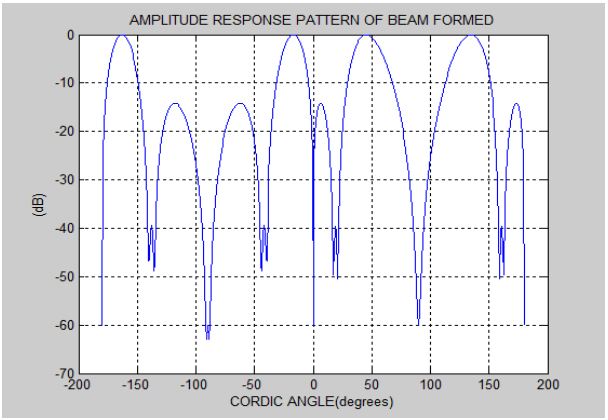


Fig.10 Output Data

As shown in fig 10 the Amplitude Response of Beamformed Data based on Cordic Angles, output is formed..

4.2. Results of Optimized Method Using MATLAB

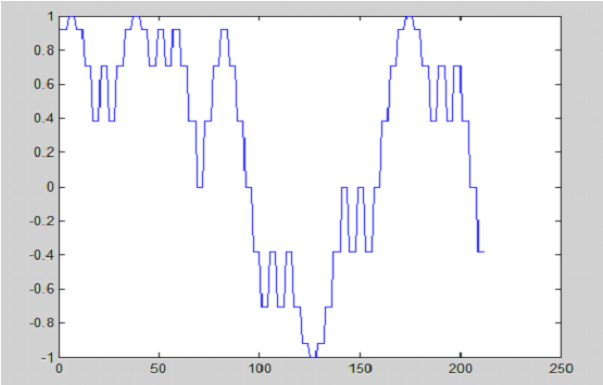


Fig.11. Phase Graph of Input Data

Fig.11. shows the input pulse sent from the transmitter side to underwater to detect the target.

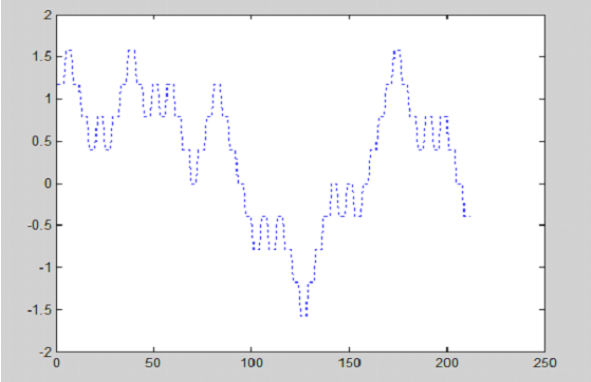


Fig.12.Beamformed Data.

Fig.12. shows the input data is sent in the form of samples from the transmitter side.

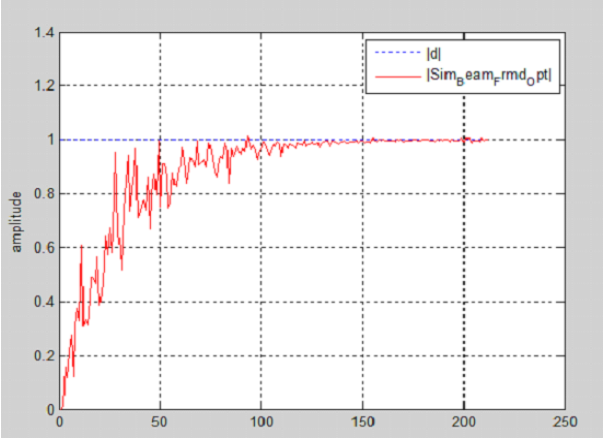


Fig.13.Amplitude Response Data

Fig.13. shows the amplitude response of optimized data from the transmitter to receiver.

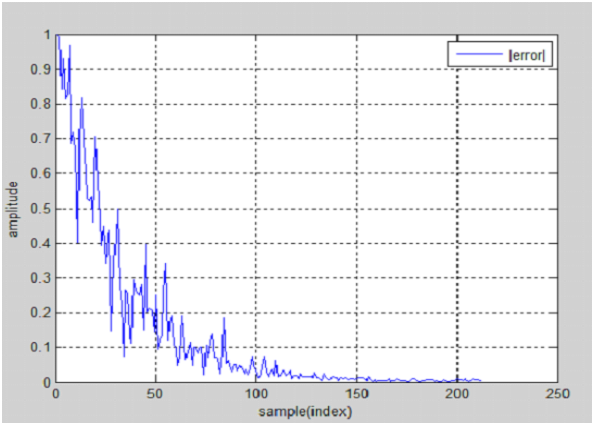


Fig.14.Error In beamformed Transmitted Data

Fig.14.shows the loss of data which is less compared to direct method.

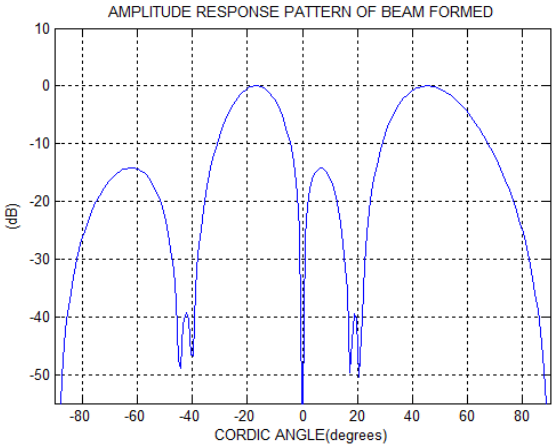


Fig.15.Amplitude Response of Beamformed Data Based on Cordic Angles

Fig.15. shows the amplitude response of beamformed data. At the transmitter side , the signal is up sampled and at the receiver side the signal is down sampled using cordic algorithm to get accurate result.

5. COMPARISON OF OPTIMIZED METHOD AND DIRECT METHOD:

Table 1. Comparison of direct method and optimized method

Parameter	Direct Method	Optimized Method	Memory reduced Optimized Method
Number of Delays per focusing distance	10^7 bytes	10^3 bytes	59kB
Validity range.	Does not enlarge.	Enlarged by 4 degree in azimuth and elevation angle	Enlarged by 4 degree
Computational requirement	More number of sensors	Less number of sensors.	Reduced by a factor of 2.

6. CONCLUSION:

This paper has illustrated that the proposed approximation enlarges the validity region of the system's view scene. Under the preferred definition of steering direction condition, the validity region is enlarged at least by 4◦ in both azimuth and elevation angles. The optimized algorithm has the advantage of reducing the memory and computational requirements as compared with DM beamforming. In high-resolution sonar systems, where more than ten thousands of beams are produced, the required memory for parameter storage is reduced.

Digital antennas have the potential of satisfying the requirements of many systems simultaneously. They are flexible, and capable of handling wide bandwidths, and can perform multiple functions. The bandwidth of the modulator and demodulator must match the bandwidth of the signal for efficient operation. The effects of the phase slope and amplitude variations on the pattern of a linear array were determined by simulations that incorporated the measured data. The simulation showed unacceptable beam squint with frequency.

7. REFERENCES:

[1] V. Murino and A.Trucco , "Three-dimensional image generation and processing in under acoustic vision," vol88,n0.12 dec 2000

[2] A.Davis and A.Lugsdin, "High speed underwater inspection for port and harbour security using coda Echoscope 3D sonar," 2005,pp.2006-2011 .

[3] R.K.Hansen and P.A Andersen, "The application of real time 3D acoustical imaging," OCEANS1998 pp.738-741

[4] M. Palmese and A. Trucco, "Digital near field beamforming for efficient 3-D underwater acoustic image generation," in Proc. IEEE Int. Workshop Imaging Syst. Tech., 2007, pp. 1-5.

[5] M. Palmese and A. Trucco, "From 3-D sonar images to augmented reality models for objects buried on the seafloor," IEEE Trans. Instrum. Meas., vol. 57, no. 4, pp. 820-828, Apr. 2008.

[6] M. Palmese, G. De Toni, and A. Trucco, "3-D underwater acoustic imaging by an efficient frequency domain beamforming," in Proc. IEEE Int. Workshop Imagining Syst. Tech., 2006, pp. 86-90.

[7] B.E. Nelson, "Configurable computing and sonar processing-arhitecture and implementations", 2001,pp.56-60

[8] B. L. Hutchings and B. E. Nelson, "Gigaop DSP on FPGA", in Proc. IEEE. Int. Conf. Acoust., Speech, Signal Process., 2001, pp. 885-888.

[9] G. Hampson and A. Paplinski, "Phase shift beamforming using cordic", in Proc. Int. Symp. Signal Process. Appl., 1996, pp. 684-687.

[10] A. Trucco, "A least-squares approximation for the delays used in focused beamforming", J. Acoust. Soc. Amer., vol. 104, no. 1, pp. 171-175, Jul. 1998.

[11] J. E. Volder, "The CORDIC trigonometric computing technique", IRETrans. Electron. Comput., vol. EC-8, no. 3, pp. 330–334, Sep. 1959

[12] J.S.Walther, "A unified algortithm for elementary funcations", in proc spring joint comput 1971,pp.379-385

[13] A. Trucco, "Enlarging the scanning region of a focused beamforming system", Electron. Lett., vol. 33, no. 17, pp. 1502-1504, Aug. 1997

[14] B. O. Odelowo, "A fast beamforming algorithm for planar/volumetric arrays", in Proc. 39th Asilomar Conf. Signals, Syst. Comput., 2005, pp. 1707-1710M.

[15] Palmese and A. Trucco, "Acoustic imaging of underwater embedded objects: Signal simulation for three-dimensional sonar instrumentation", IEEE Trans. Instrum. Meas., vol. 55, no. 4, pp. 1339–1347, Aug. 2006.

.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 822 - 826, 2014, ISSN:- 2319–8656

Local Restoration in Metro Ethernet Networks for Multiple Link Failures

Shibu. V
Department of Computer Applications.
Cochin University College of Engineering, Pulincunnoo.
Alappuzha, Kerala, India

Preetha Mathew K
Department of Computer Applications.
Cochin University College of Engineering, Pulincunnoo.
Alappuzha, Kerala, India

Jabir.K.V.T
Department of Information Technology.
Cochin University College of Engineering, Pulincunnoo.
Alappuzha, Kerala, India

Abstract: Ethernet a popular choice for metropolitan-area networks (MAN) due to simplicity, cost effectiveness and scalability. The Spanning-Tree based switching mechanism, which is considered to be very efficient at avoiding switching loops in LAN environment, is a performance bottleneck in Metro network context. Handling of link failure is an important issue in metro Ethernet networks. A link failure may result in serious service disruptions. A local restoration method for metro Ethernet with multiple spanning trees, which aims at fast handling of single link failures in a distributed manner, have been proposed in the literature. In this paper, we propose a local restoration mechanism that uses MULTILINK algorithm for solving multiple link failures

Keywords: Metropolitan Area Networks (MAN), Ethernet, Spanning Tree Protocol, RSTP.

1. INTRODUCTION

Ethernet is a family of computer networking technologies for local area networks (LANs). Systems communicating over Ethernet divide a stream of data into individual packets called frames. Each frame contains source and destination addresses and error-checking data so that damaged data can be detected and re-transmitted. Ethernet has evolved over the past decade from a simple shared medium access protocol to a full-duplex switched network. Ethernet dominates current local area network (LAN) realizations. It has been estimated that more than 90 percent of IP traffic originates from Ethernet LANs. Efforts are underway to make Ethernet an end-to-end technology spanning across LANs, metropolitan area networks (MANs), and possibly wide area networks (WANs) [2].

A Metro Ethernet [1] is a computer network that covers a metropolitan area and that is based on the Ethernet standard. It is commonly used as a metropolitan access network to connect subscribers and businesses to a larger service network or the Internet. Metro Ethernet network is a set of interconnected LANs and access networks that work together using Ethernet technologies to provide access and services within a metro region. Metro Ethernet networks are built from Ethernet switches/ bridges interconnected by fiber links. A spanning tree protocol is used to establish one or more trees spanning every access point that connects LANs.

Failure handling is a key issue in metro Ethernet networks. A component failure may result in serious service disruptions. To support carrier-grade services in MANs using Ethernet, it is a critical requirement to have a fast, reliable, and efficient failure-handling mechanism [3]. Current Ethernet switched networks use the spanning tree protocol family without any fast failure recovery mechanism. The IEEE 802.1d Spanning Tree Protocol (STP) [4] establishes a single spanning tree to guarantee a unique path between any two switches. It suffers from low convergence speed and inefficient bandwidth usage in case of a failure.

The spanning tree approach fails to exploit all the physical network resources, because in any network of N nodes there are at most N−1 links actively forwarding traffic. This produces an imbalance of load in the network. This scenario is impractical in large scale networks like metro networks. Further, switch and link failures require rebuilding of the spanning tree, which is a lengthy process. IEEE 802.1w [5], the rapid spanning tree configuration protocol (RSTP), mitigates this problem by providing mechanisms to detect failures and quickly reconfigure the spanning tree. However, the recovery period can still range from an optimistic 10 milliseconds to more realistic multiple seconds *after* failure detection, which is still not adequate for many applications.

In this paper, we propose a local restoration mechanism for metro Ethernet, which aims at fast handling of multiple link failures in an efficient manner. Multiple link failures come from the fact that when a connection is switched to a backup spanning tree, it has no record of the original working spanning tree. Therefore, when the connection encounters another failure on the backup spanning tree, there is a possibility that it would be switched back to the original working spanning tree and form a loop in the network. However, when multiple link failure happens in the network and both the primary and backup spanning tree fail simultaneously, some packets would be dropped when they encounter the failure on the backup spanning tree. We propose a possible approach to handle these multiple link failures. This approach is to allow multiple VLAN switching and add more information in the header of the frames, e.g., VLAN ID of the original working spanning tree, when the frames are switched to a backup tree. Thus, they are able to select a backup tree without forming a loop when they are affected by the second failure in the network.

2. PROBLEM DEFINITION

The simplicity and the low cost provided by Ethernet makes it an attractive network technology choice in networking application deployments. The Spanning-Tree Protocol (STP), which is proposed in initial version of IEEE 802.1D [4], is responsible for building a loop-free logical forwarding topology over the physical one providing connectivity among all nodes. The links that are not part of this tree are blocked. In case of a failure, the blocked links are activated providing a self-healing restoration mechanism. All information propagated between the switches is embedded in Bridge Protocol Data Units (BPDUs). These packets are exchanged only between adjacent bridges, and protocol events (e.g., port state changes) are invoked by timers, thus rebuilding the topology takes considerable time. This timer based operation, which is an STP property, results in reconfiguration times up to 50 seconds and, thus, affects network performance. The existing system defines a local restoration mechanism for metro Ethernet using multiple spanning trees, which is distributed and fast and does not need failure notification. Upon failure of a single link, the upstream switch locally restores traffic to preconfigured backup spanning trees. There are two restoration approaches, connection-based and destination-based, to select backup trees. When multiple link failure happens in the network and both the primary and backup spanning tree fail simultaneously, some packets would be dropped when they encounter the failure on the backup spanning tree. The proposed system defines a possible approach to handle these multiple link failures. This approach is to allow multiple VLAN switching and add more information in the header of the frames, e.g., VLAN ID of the original working spanning tree, when the frames are switched to a backup tree. Thus, they are able to select a backup tree without forming a loop when they are affected by the second failure in the network.

3. EXISTING METHODS

3.1 Metro Ethernet Local Restoration Framework

The existing system define a local restoration mechanism[1] in metro Ethernet that selects appropriate backup spanning trees for rerouting traffic on a working spanning tree. Then restores the traffic to the backup spanning trees in case of failure locally. The path on a backup tree to reroute traffic is from the immediate upstream node of the failed link to the destination node and should exclude the failed link. Using the current Ethernet protocols the local restoration mechanism in metro Ethernet can be implemented. For this, an additional module should be maintained in the Ethernet switch for checking the appropriate backup spanning tree and restoring frames to the backup tree after failure.

The existing system has implemented the local restoration mechanism in the following three steps.

3.1.1 Per VLAN Spanning Tree:

Local restoration from one spanning tree to another can be implemented by assuming in each spanning tree is assigned a dedicated virtual LAN (VLAN) ID [6]. The pre-calculated spanning tree topologies are implemented in the network by means of VLANs, which do not change during network operation and ensure that there are no loops in the Ethernet network. Therefore, STP is disabled, as it is not needed to provide loop free topology. A unique VLAN ID is assigned to each spanning tree, which is used by the edge routers to forward traffic over the appropriate trees [3]. By changing the VLAN ID on the Ethernet header, Ethernet frames can be switched among spanning trees. Frames that frequently switch among spanning trees may form unexpected loops. To allow VLAN switching once, set a bit in the frame's class-of-service (CoS) field as the restoration bit. The Ethernet switch will check a frames restoration bit before restoration and drop those frames that have been earlier restored once.

3.1.2 Local Restoration Mechanism:

Local restoration doesn't require the convergence of spanning trees after failure. To inform that the switch is alive, each switch periodically sends a message to its neighbors. Within a predefined interval, if a switch does not receive any message from a port, it changes the port's status as "failed". The restoration module is activated when an incoming frame is forwarded to the failed port; a preconfigured backup VLAN ID replaces the frame's original VLAN ID. At the same time, its restoration bit is set to 1. Then, the modified frame is forwarded to the alternative output port according to its new VLAN ID.

3.1.3 Local Restoration Mechanism:

The network manager will perform the pre configuration operation which includes three parts: multiple spanning trees generation, working spanning tree assignment, and backup spanning tree configuration.

Multiple Spanning Trees Generation: The network manager is responsible for generating multiple spanning trees [8] *a priori*. The trees should satisfy the condition to handle a link failure: For each link, there is at least one spanning tree that does not include that particular link [7]. Commonly, more spanning trees should be generated to utilize network resources efficiently.

Working Spanning Tree Assignment: The network manager should assign a VLAN ID to each source and destination (s–d) pair based on the long-term traffic demand matrix. The frames entering the network are attached with VLAN IDs at the ingress switches according to their source and destination addresses and are forwarded to the proper working spanning trees.

Backup Spanning Tree Configuration: Frames traversing the failed link should be restored to a preconfigured backup spanning tree locally when a failure happens. Backup trees at each switch should be carefully configured according to the traffic demand such that there is enough spare capacity on the backup tree for restoration.

Backup Tree Selection Strategy

The Ethernet switch selects the backup tree for each frame according to the backup tree configuration by the network manager. The existing system uses two backup spanning tree selection strategies: connection-based and destination-based. The traffic between a source–destination pair is termed as a connection. In connection-based strategy an Ethernet switch determines the incoming frame's backup VLAN ID according to its source address, destination address, and original VLAN ID. Therefore, traffic between different source–destination pairs traversing the same failed link may be restored to different backup trees

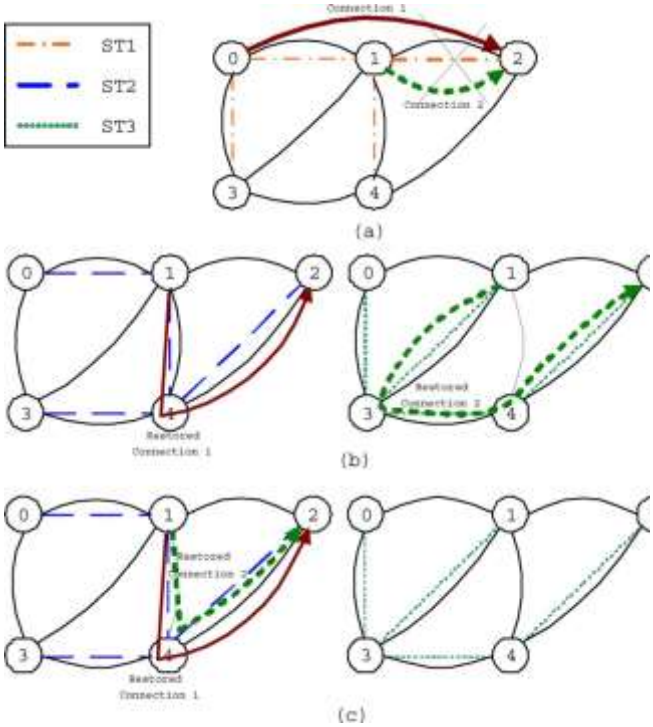


Fig.1 Backup tree selection strategy. (a) Two connections on ST1 before failure of link 1–2. (b) Two connections are restored to different STs after failure in connection-based strategy. (c) Two connections are restored to the same ST in destination-based strategy.

Connection 1 from node 0 to 2 and connection 2 from node 1 to 2 uses ST1 as the working spanning tree before failure [Fig. 1(a)]. When the link between 1–2 fails, node 1 restores connection 1 to ST2 and connection 2 to ST3 according to the pre configuration. According to their source and destination MAC addresses frames are restored and different connections are assigned independent backup spanning trees. Connection-based backup tree selection requires a complex computation during pre configuration and per-(source–destination) pair information should be maintained by each switch.

The existing system uses a destination- based backup tree selection strategy, in which the frame's backup VLAN ID is determined by its destination address and original VLAN ID, regardless of its source address. Frames with the same VLAN ID and destination address would use the same backup tree in a local Ethernet switch. Fig. 1(c) shows an example of the destination-based backup tree selection strategy. Connections 1 and 2 have to use the same backup spanning tree in node 1 upon failure of link 1–2 since they have the same destination which is different from the connection-based strategy. Node 1 can only restore the two connections with the same destination to the same spanning tree.

4. PROPOSED METHODOLOGY

When a connection is switched to a backup spanning tree and when the connection encounters another failure on the backup spanning tree, multiple link failure happens in the network. Both the primary and backup spanning tree fail simultaneously, some packets would be dropped when they encounter the failure on the backup spanning tree.

The existing system only handles single link failures in the metro Ethernet. The proposed system is an enhancement of the existing system. This system defines a possible approach to handle these multiple link failures. The approach is to allow multiple VLAN switching and add more information in the header of the frames, e.g., VLAN ID of the original working spanning tree, when the frames are switched to a backup tree. Thus, they are able to select a backup tree without forming a loop when they are affected by the second failure in the network. It also uses the concept of local restoration mechanism provided in the existing system in its restoration module. The restoration module uses an algorithm namely MULTILINK. The working of this proposed algorithm is demonstrated in the next section.

The MULTILINK algorithm

This algorithm considers three connections. We named these connections as TLink, Blink and RLink respectively. In the first case consider all the three links are up. Two bits in the Ethernet IP packet header are used in this algorithm. The first is the restoration bit (RB) that is used in the existing system. The second bit is termed as new bit (NB) in the proposed restoration module, which will be assigned by a value 1 or 0 according to various conditions in the algorithm. After considering that all the three links are up, assign value zero to RB and NB in the second step. If the TLink is down then set the value of RB to one and set the value of NB to zero. Then switch to the Blink by activating the restoration module as in the existing system. Now the traffic is going through the Blink. Check whether the TLink is up frequently within a fixed time interval of 5Ms. If the TLink is up then set the value of the bit NB to one. Otherwise set the value of the bit NB to zero. If the Blink is up, repeat the said operations by checking TLink is down or not. If the Blink is down then check whether the value of NB. If the value of NB is equal to one then route the packet through TLink otherwise route the packet through RLink by using Rapid Spanning Tree Protocol (RSTP). The working of MULTILINK algorithm is demonstrated in the following Figures.

5. EXPECTED RESULTS

The most important design objectives of a failure handling mechanism are fast failover, simplicity, robustness and low protocol processing and transport overheads. Ethernet has built-in functionalities for failure handling developed in standardization bodies. When a connection is switched to a backup spanning tree and when the connection encounters another failure on the backup spanning tree, multiple link failure happens in the network. Both the primary and backup spanning tree fail simultaneously, some packets would be dropped when they encounter the failure on the backup spanning tree. When an Ethernet switch finds a packet that has been rerouted once and its output port on the backup tree also fails, the switch should notify the network manager or broadcast the failure message asking for spanning tree re convergence by RSTP. The algorithm invokes RSTP only when the last condition NB is equal to one doesn't satisfy. The MULTILINK algorithm described in the proposed system, on implementation, can solve the multiple link failure efficiently

Algorithm is defined as follows,

Algorithm MULTILINK

1. Initially all the three links TLink, Blink and RLink are considered as in up state.
2. If TLink is down then
   - A) set RB=0 and NB=0
   - B) Switch to Blink by activating the restoration module.
   - C) Check whether TLink is up frequently by setting a time limit of 5ms.
     - (i) if TLink is up then

       set NB=1

       else
       - (a) set NB=0
       - (b) if Blink is down then go to step 3

       else go to step 2.
3. If NB=1 then

   Route the packet through TLink

   Else

   Route the packet through RLink.

Fig.2 Algorithm

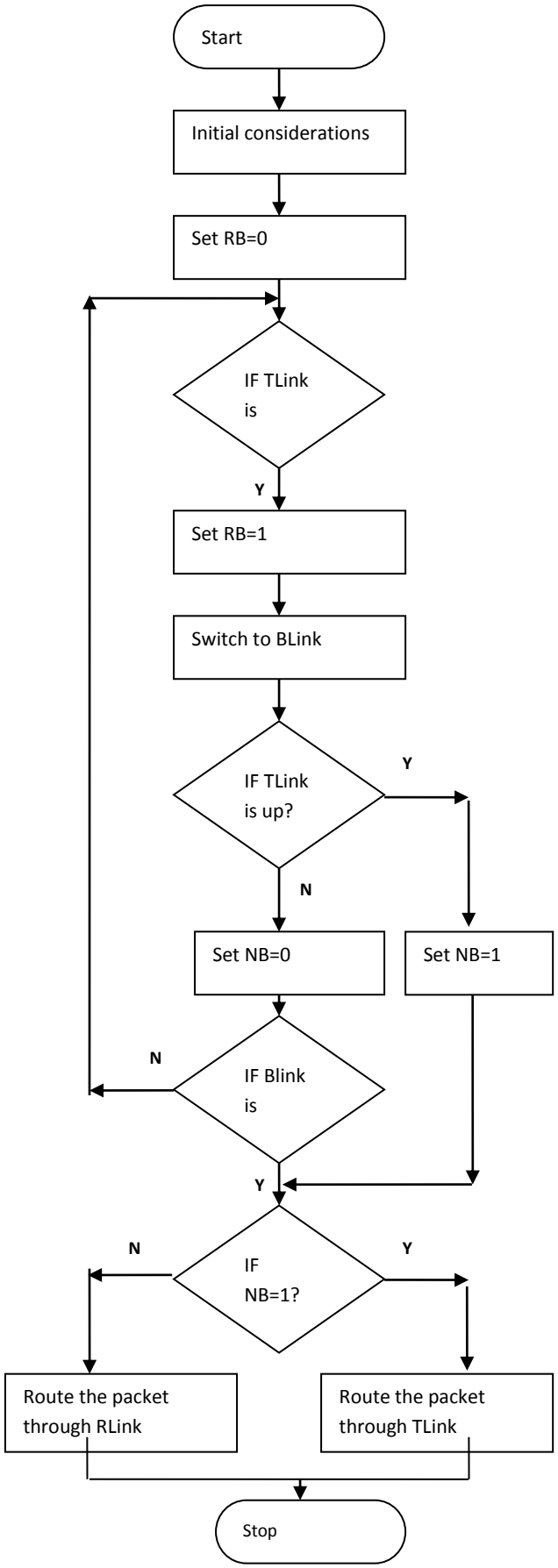


Fig.3 Flowchart working of multi link algorithm

6. CONCLUSION

The existing system only handles single link failures in the metro Ethernet. The proposed system is an enhancement of the existing system. This system defines a possible approach to handle these multiple link failures. It also uses the concept of local restoration mechanism provided in the existing system in its restoration module. The restoration module uses an algorithm namely MULTILINK. Two bits in the Ethernet IP packet header are used in this algorithm. The occurrence of multiple link failure is a rare event. When implemented properly, the proposed system solves the problem of multiple link failures in metro Ethernet network.

7. REFERENCES

[1] Jian Qiu, Mohan Gurusamy, "Local Restoration With Multiple Spanning Trees in Metro Ethernet Networks", IEEE/ACM Transactions On Networking, Vol. 19, No. 2, April 2011

[2] A. Meddeb, "Why Ethernet WAN transport," *IEEE Commun. Mag.*,vol. 43, no. 11, pp. 136–141, Nov. 2005.

[3] C. Antal, L. Westberg, A. Paradisi, T. R. Tronco, and V. G. Oliveira, "Fast failure handling in Ethernet network," in *Proc IEEE ICC*, 2006, vol. 2, pp. 841–846.

[4] Standard for Local and Metropolitan Area Networks—Media Access Control (MAC) Bridges, IEEE 802.1d, 1998.

[5] Standard for Local and Metropolitan Area Networks—Rapid Reconfiguration of Spanning Tree, IEEE 802.1w, 2001.

[6] Standard for Local and Metropolitan Area Networks—Virtual Bridged Local Area Networks, IEEE 802.1q, 1999.

[7] J. Farkas, C. Antal, G. Toth, and L.Westberg, "Distributed resilient architecture for Ethernet networks," in *Proc. DRCN*, 2005, pp. 515–522.

[8] K. Goplan, S. Nanda, and T. Chiueh, "Viking: A multiple-spanning tree Ethernet architecture for metropolitan area and cluster networks," in *Proc. IEEE INFOCOM*, 2004, pp. 2283–2294.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 827 - 830, 2014, ISSN:- 2319–8656

Prediction Model Using Web Usage Mining Techniques

Priyanka Bhart
U.I.E.T Kurukshetra University
Kurukshetra, India

Abstract: Popularity of WWW increasing day by day which results in the increase of web based services , due to which web is now largest data repository. In order to handle this incremental nature of data various prediction techniques are used. If the prefetched pages are not visited by user in their subsequent access there will be wastage network bandwidth as it is in limited amount. So there is critical requirement of accurate prediction method. As the data present on web is heterogeneous in nature and incremental in nature, during the pre-processing step hierarchical clustering technique is used. Then using Markov model category and page prediction is done and lastly page filtering is done using keywords.

Keywords: formatting Hierarchal clustering; markov model; page prediction; category prediction ;

1. INTRODUCTION

With the continued growth and proliferation of E-commerce there is need to predict users behavior. These predictions helps in implementing personalization, building proper websites, improving marketing strategy promotion, getting marketing information, forecasting market trends, and increase the competitive strength of enterprises etc.[1].

Web prediction is one of the classification problem where a set of web pages a user may visit are predicted on the basis of previously visited page which are stored in the form of web log files. Such kind of knowledge of users' navigation history within a slot of time is referred to as a session. This data is extracted from the log files of the web server which contains the sequence of web pages that a user visits along with visit date and time. This data is fed as the training data.

All the user's browsing behavior is recorded in the web log file with user's name, IP address, date, and request time etc.

S.no	Ip address	Req.	Timestamp	Protocol	Total bytes

Table 1.Common web log

Hierarchical clustering is a classification technique. It is an agglomerative(top down) clustering method , as its name suggests, the idea of this method is to build hierarchy of clusters, showing relations between the individual members and merging clusters of data based on similarity.

In order to analyze user web navigation data Markov model is used. It is used in category and page prediction. Only the set of pages which belong to the category which is predicted in first phase are used in second phase of page prediction. Here each Web page represent a state and ever pair of pages viewed in sequence represent a state transition in this model. The transition probability is calculated by the ratio of number of a particular transition is visited to the number of times the first state in the pair was visited.

This paper introduces an efficient four stage prediction model in order to analyze Web user navigation behavior .This model is used in identification of navigation patterns of users and to anticipate next choice of link of a user. It is expected that that this prediction model will reduce the operation scope and increase the accuracy precision.

2. RELATED WORK

Agrawal R and Srikant R [1] proposed a website access prediction method based on past access behavior of user by constructing first-order and second-order Markov model of website access and compare it association rules technique. Here by using session identification technique sequence of user requests are collected, which distinguishes the requests for the same web page in different browses. Trilok Nath Pandey [2] proposed a Integrating Markov Model with Clustering approach for user future request prediction. Here improvement of Markov model accuracy is done by grouping web sessions into clusters. The web pages in the user sessions are first allocated into categories according to the web services that are functionally meaning full. And lastly k-means clustering algorithm is implemented using the most appropriate number of clusters and distance measures. Lee and Fu proposed a two level prediction model in 2008[4]. Here prediction scope is reduced as it works in two levels. This model is designed by combining Markov Model and Bayesian theorem. In level one using Markov Model category prediction is done and page prediction is done by Bayesian theorem. Chu-Hui Lee [3] used the hierarchical agglomerative clustering to cluster user browsing behaviors due to the heterogeneity of user browsing features. The prediction results by two levels of prediction model framework work well in general cases. However, two levels of prediction model suffer from the heterogeneity user's behavior. So they have proposed a prediction model which decreases the prediction scope using two levels of framework. This prediction model is designed by combining Markov model and Bayesian theorem.

Sujatha [5] proposed the prediction of user navigation patterns using clustering and classification (PUCC) from web log data. In the first phase it separates the potential users in web log data, and in second t-stage clustering process is used to group the potential users with similar interest and lastly the results of classification and clustering is used to predict the users future requests. Sonal vishwakarma [6] analyzed all order Markov model with webpage keywords as a feature to give more accurate results in Web prediction.

3. OVERVIEW OF THE PROPOSED PREDICTION FRAMEWORK

The prediction model is designed by combining clustering and Markov model technique. During preprocessing step hierarchical clustering is done to group user's browsing behaviors and acquires many different clusters. The information of relevant to any cluster can be seen as cluster view that means every cluster has its own relevant matrix irrespective of having these matrixes for every user, so here global view is replaced by cluster view. After preprocessing category prediction is done by using Markov model. Here in this phase it is to predict category at time t which depends upon users category time at time t-1 and t-2. In the same way page prediction is done to predict the most possible web pages at a time t according to users state at a time t-1. Now the set of predicted pages are fed for keyword based filtering. Finally after this phase predicted results are released.

Firstly training data is fed for clustering where k number of cluster view will be obtained which include k similarity matrices S, k first-order transition matrices P and k second-order transition matrices between categories. Therefore, we get K relevant matrices R to represent K cluster views [4]. In step two, these matrices will be released out for creating index table which will be used for view selection based on user's browsing behavior at that time. In step three, after view selection testing data is fed into the prediction model and prediction results will be released as output.

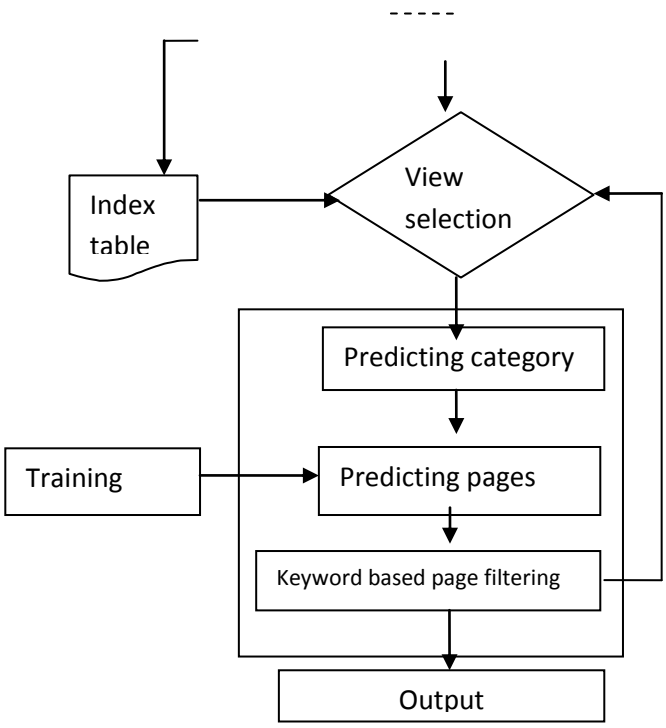
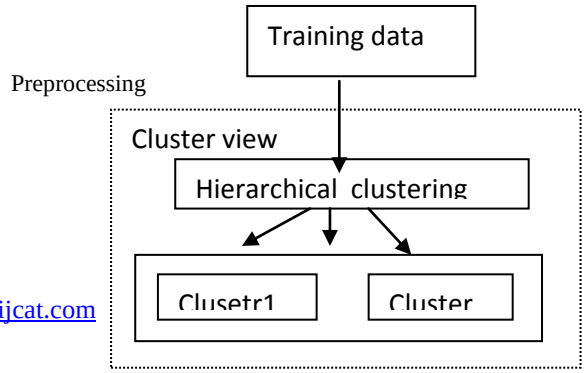


Figure 1. Proposed prediction framework

3.1 HIERARCHICAL CLUSTERING ALGORITHM

It is an agglomerative clustering method. The idea of this method is to build a hierarchy of clusters, showing relations between members and merging clusters of data based on similarity. For the clustering algorithm to work there is need to have some means by which similarity to be judged. This is generally called a distance measurement. Two commonly used metrics for measuring correlation (similarity/distance) are Euclidean and the Pearson correlations. The type of correlation metric used depends largely on what it is to be measured. Here we have used Euclidean correlation.

In the first step of clustering, the algorithm will look for the two most similar data points and merge them to create a new "pseudo-data point", which represents the average of the two data points. Each iterative step takes the next two closest data points and merges them. This process is generally continued until there is one large cluster covering all original data points. This clustering technique will results in a "tree", showing the relationship of all the original points. Here every user seems to be a cluster and grouped by most similar browsing feature into the cluster [12].



3.2 MARKOV MODEL

Markov is a probability based model which is represented by three parameters<A,S,T> where A is a set of all possible actions performed by any user; S is the set all possible states for which the Markov model is built; and T is a |A|× |S| Transition probability matrix, where represent the probability of performing the action j when the process is in state i. Markov model predicts the user's next action by looking at previously performed action by the user.

Here assume that D is the given database, which consists of user's usage records. It means users sessions are recorded and $D=\{session_1, session_2, \ldots\ldots\ldots, session_p\}$. each user session is a set of web pages recorded in time orderd sequential pattern and $session_p=\{page_1, page_2, \ldots\ldots, page_n\}$, where $page_i$ represents user's visiting page at time j. If a website has K categories, then the user session can be represented as $session_c=\{c_1, c_2, \ldots.., c_k\}$.

3.3 TRANSITION MATRIX

P is transition probability matrix represented as in equation(1) where P_{ij} represent the transition probability between any two pages/category ie. from P_i to P_j. It is calculated by the ratio of number of transition between category/page i and category/page j to the total number of transition between category/page i and every category/page k.

$$P = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1k} \\ P_{21} & P_{22} & \cdots & P_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k1} & P_{k2} & \cdots & P_{kk} \end{bmatrix} \end{array} \quad (1)$$

Web sessions

WS1={P3,P2,P1}

WS2={P3,P5,P2,P1,P4}

WS3={P4,P5,P2,P1,P5,P4}

WS4={P3,P4,P5,P2,P1}

WS5={P1,P4,P2,P5,P4}

1st Order	P1	P2	P3	P4	P5
S1={P1}	0	0	0	2	1
S2={P2}	4	0	0	0	1
S3={P3}	0	1	0	1	1
S4={P4}	0	1	0	0	2
S5={P5}	0	3	0	2	0

2nd Order	P1	P2	P3	P4	P5
{P1,P4}	0	1	0	0	0
{P1,P5}	0	0	0	1	0
{P2,P1}	0	0	0	1	1
{P2,P5}	0	0	0	1	0
{P3,P2}	1	0	0	0	0

2nd Order	P1	P2	P3	P4	P5
{P2,P5}	0	1	0	0	0
{P2,P4}	0	0	0	0	1
{P4,P5}	0	2	0	0	0
{P5,P2}	3	0	0	0	0
{P3,P4}	0	0	0	0	1

Figure 2. Sample web sessions with corresponding 1_{st} and 2^{nd} order transition probability matrices [7].

3.4 SIMILARITY MATRIX

Similarity between any two user user i and user j, can be calculated using Euclidean distance given in equation (3) .

$$sim(user_i, user_j) = (session^i, session^j) \quad \textbf{(2)}$$

Euclidean distance

$$D(user_i, user_j) = \sqrt{\sum_{i=1}^{k} (P_{i,i} - P_{j,l})^2} \quad (3)$$

Euclidean distance is further normalized by (4) equation, by this k×k similarity matrix as given in equation (5) will be obtained.

$$ND(user_i, user_j) = 1 - \sqrt{\frac{\sum_{i=1}^{k} (P_{i,i} - P_{j,l})^2}{k}} \quad (4)$$

$$S = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1k} \\ S_{21} & S_{22} & \cdots & S_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k1} & S_{k2} & \cdots & S_{kk} \end{bmatrix} \end{array} \quad (5)$$

3.5 RELEVANCE MATRIX

Last matrix to create is relevance matrix represented as in equation (6),which is equal to the product of transition and similarity matrix. Here relevance is an important factor of prediction between any two category and pages. It concludes the behavior between pages and categories. It is represented as follows:

$$R_n = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} {R_{11}}^n & {R_{12}}^n & \cdots & {R_{1k}}^n \\ {R_{21}}^n & {R_{21}}^n & \cdots & {R_{2k}}^n \\ \vdots & \vdots & \ddots & \vdots \\ {R_{k1}}^n & {R_{k2}}^n & \cdots & {R_{kk}}^n \end{bmatrix} \end{array} \quad (6)$$

Where

$${R_{ij}}^n = {P_{ij}}^n \times S_{ij} \quad (7)$$

4. CONCLUSION

As there is large amount of data on web pages on many websites, So it is better to place them according to their category. In this paper users browsing behavior is firstly preprocessed using hierarchical clustering then prediction is done in three phases. In first phase category prediction is done using Markov model then in second phase page prediction is done. And lastly keyword based filtering is done which gives more accurate results.

5. REFERENCES

.

[1] Agrawal R, Imielinski T and Swami A "Mining Association Rules between Sets of Items in Large Databases", ACM SIGMOD Conference on Management of Data, pp.207-216.

[2] Trilok Nath Pandey, Ranjita Kumari Dash , Alaka Nanda Tripathy , Barnali Sahu, "Merging Data Mining Techniques for Web Page Access Prediction: Integrating Markov Model with Clustering", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 1, November 2012.

[3] Chu-Hui Lee, Yu-Hsiang Fu "Web Usage Mining based on Clustering of Browsing Features**"** Eighth International Conference on Intelligent Systems Design and Applications,IEEE, 2008. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[4] Chu-Hui Lee , Yu-lung Lo, Yu-Hsiang Fu, "A Novel Prediction Model based on Hierarchical Characteristic of Web Site", Expert Systems with Applications 38 , 2011.

[5] V. Sujatha, Punithavalli, "Improved User Navigation Pattern Prediction Technique From Web Log Data", Procedia Engineering 30 ,2012.

[6] Sonal Vishwakarma, Shrikant Lade, Manish Kumar Suman and Deepak Patel "Web User Prediction by: Integrating Markov with Different Features",vol2 IJERST , 2013.

[7] Deshpande M and Karypis G (2004), "Selective Markov Models for Predicting Web-Page Accesses", ACM Transactions on Internet Technology (TIOIT),Vol.4,No.2,pp.163-184.

[8] UCI KDD archive, http://kdd.ics.uci.edu/

[9] V.V.R. Maheswara Rao, Dr. V. Valli Kumari" An Efficient Hybrid Predictive Model to Analyze the Visiting Characteristics of Web User using Web Usage Mining" 2010 International Conference on Advances in Recent Technologies in Communication and Computing IEEE.

[10] A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction", International Journal of Computer Applications, Volume8–No.11,October2010.

[11] Mehrdad Jalali, Norwati Mustapha, Md. Nasir Sulaiman, Ali Mamat, "WebPUM: A Web-Based Recommendation System to Predict User Future Movements" Expert Systems with Applications 37 , 2010.

[12] www.microarrays.ca/services/hierarchical_clustering.pdf

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 831 - 835, 2014, ISSN:- 2319–8656

Performance Prediction of Service-Oriented Architecture - A survey

Haitham A.Moniem,
College of Graduate Studies,
Sudan University of Science and Technology,
Khartoum, Sudan

Hany H Ammar,
Lane Department of Computer Science and Electrical Engineering,
College of Engineering and Mineral Resources,
West Virginia University
Morgantown, USA

Abstract: Performance prediction and evaluation for SOA based applications assist software consumers to estimate their applications based on service specifications created by service developers. Incorporating traditional performance models such as Stochastic Petri Nets, Queuing Networks, and Simulation present drawbacks of SOA based applications due to special characteristics of SOA such as lose coupling, self-contained and interoperability. Although, researchers have suggested many methods in this area during last decade, none of them has obtained popular industrial use. Based on this, we have conducted a comprehensive survey on these methods to estimate their applicability. This survey classified these approaches according to their performance metrics analyzed, performance models used, and applicable project stage. Our survey helps SOA architects to select the appropriate approach based on target performance metric and researchers to identify the SOA state-of-art performance prediction.

Keywords: Service; Service-Oriented Architecture; Performance; Prediction; Evaluation

1. INTRODUCTION

Service-Oriented Architecture (SOA) is an architectural style as well as technology of delivering services to either users or other services through a network. SOA architecture created in order to satisfy business goals that include easy and flexible integration with other systems. SOA has many advantages such as reducing development costs, creative services to customers, and agile deployment [1].

There are many definitions for SOA, but they are all point to the same core idea that SOA is simply a collection of application services. The service defined as "a function or some processing logic or business processing that well-defined, self-contained, and does not depend on the context or state of other services" [2]. It also states that "Generally SOA can be classified into two terms: Services and Connectors."

Open Management Group (OMG) defines SOA as: "an architectural style that supports service orientation". It goes further to define service orientation, "service orientation is a way of thinking in terms of services and services-based development and the outcomes of services". Moreover, SOA is communication between services and applications which sometimes involves data transfer. But the communication between applications does not happen as a point-to-point interaction; instead it happens through a platform-independent, general purpose middle-ware that handles all communications by the use of web services [2].

The main goal of this paper is to report a detail survey in performance prediction of SOA. Section 2 lays the important concepts of SOA. Section 3 explains by a diagram an example of SOA base application and how the system exchanges the messages. Section 4 presents the performance metrics of SOA based applications. We considered three important metrics which are response time, throughput, and resource utilization. Section 5 summarizes the previous work in table 1 mentioning the date of published paper, the name of authors, objectives, performance metrics, performance model, and applicability stage. Section 5 concludes the paper.

2. SERVICE-ORIENTED ARCHITECTURE CONCEPTS

This part briefly tries to describe some important concepts related to SOA.

2.1 Enterprise Service Bus (ESB)

An ESB is a standard infrastructure that combines messaging, web services, data transformation, and intelligent routing in a highly distributed and different environment [7] [9].

2.2 Business Process Execution Language (BPEL)

BPEL is a language for designing SOA based systems. It contains a lot of facilities such as web services composition, publishing available services, organizing service execution, and handling exceptions.

2.3 ACME

ACME is a generic language for describing software architecture. It presents constructs for describing systems as graphs of components interacting through connectors [11].

3. EXAMPLE

Figure 1, present an example of SOA architecture. The example will explain in steps the requests and responses flow between service provider, service consumer, and the directory. Step 1 Service provider publishes its service description on a directory, step 2 Consumer performs queries to the directory to locate a service and find out to communicate with the provider, step 3 Service description is written in a special language called Web Service Description Language (WSDL), step 4 Messages are sent and received from the directory in a special language called Simple Object Access Protocol (SOAP), step 5 Consumer formulate its message to the provider using tag based language called Extensible Markup Language (XML). The message is generated in XML but it is

based on specifications defined in WSDL, step 6 the response generated by the provider is also in tag based XML format.

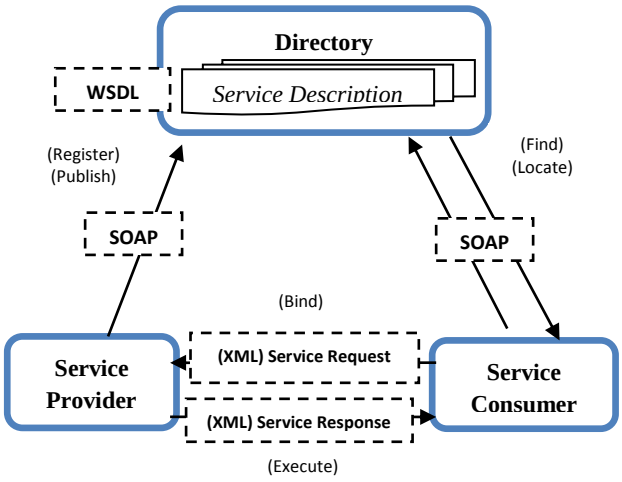


Figure. 1 Example of SOA Architecture

Based on ISO 9126 performance metrics are response time, throughput, and resource utilization [12]. Therefore, accurate measuring of SOA application plays an important role to business success. If the application has an efficient performance, this will lead to high productivity, well hardware utilization, and customer satisfaction. Otherwise, SOA based application capability will have limited benefits, resource wasting, low productivity, and unsatisfied customer.

Applying performance to SOA applications one of the challenging non-functional quality attribute, this because of the physical geographic distribution of services, communication overhead, use of standard message format, and varying service workload [3]. Performance evaluation and analysis differs in each situation of SOA based application. However, previous works on service performance are not accurate and practical enough to effectively understand and diagnose the reasons behind performance degradation.

4. SERVICE-ORIENTED ARCHITECTURE PERFORMANCE METRICS

4.1 Service Response Time

Service Response Time is the measure of the time between the end of a request to a service and the beginning of the time service provider response. There are many considerations to measure service response time [4] as Figure 2 stated. The main reasons that cause low performance of SOA based applications are:

- Services provider and service requester are positioned at different geographical areas, mostly at different machines.
- The potential problems of XML which is the standard message format increases the time needed to process a request.
- The time needed to discover the services through the directory either in design time or run time.
- Rules that govern services contain a business process by business process's need.
- Adaptation of service composition by adding new service or adapting existing services.
- Think time is an elapsed time between the end of a response time generated by a service and the beginning of an end user's request [4].

4.2 Throughput

Throughput defined as the number of requests SOA application can process at a given period of time. There are two metrics for throughput; throughput of a service and throughput of a business process [4] as Figure 3 stated.

The value range of these two metrics service throughput and business process throughput must be greater than zero. The higher the values indicate a better SOA application performance.

4.3 Resource Utilization

To analyze the performance of SOA based applications in terms of resource utilization, there are three basics information needed: firstly, workload information, which consists of concurrent users, and request arrival rates. Secondly, software specification, which consists of execution path, components to be executed, and the protocol of contention used by the software [5]. Finally, environmental information, this information consists of system specification such as configuration and device service rates, and scheduling policies.

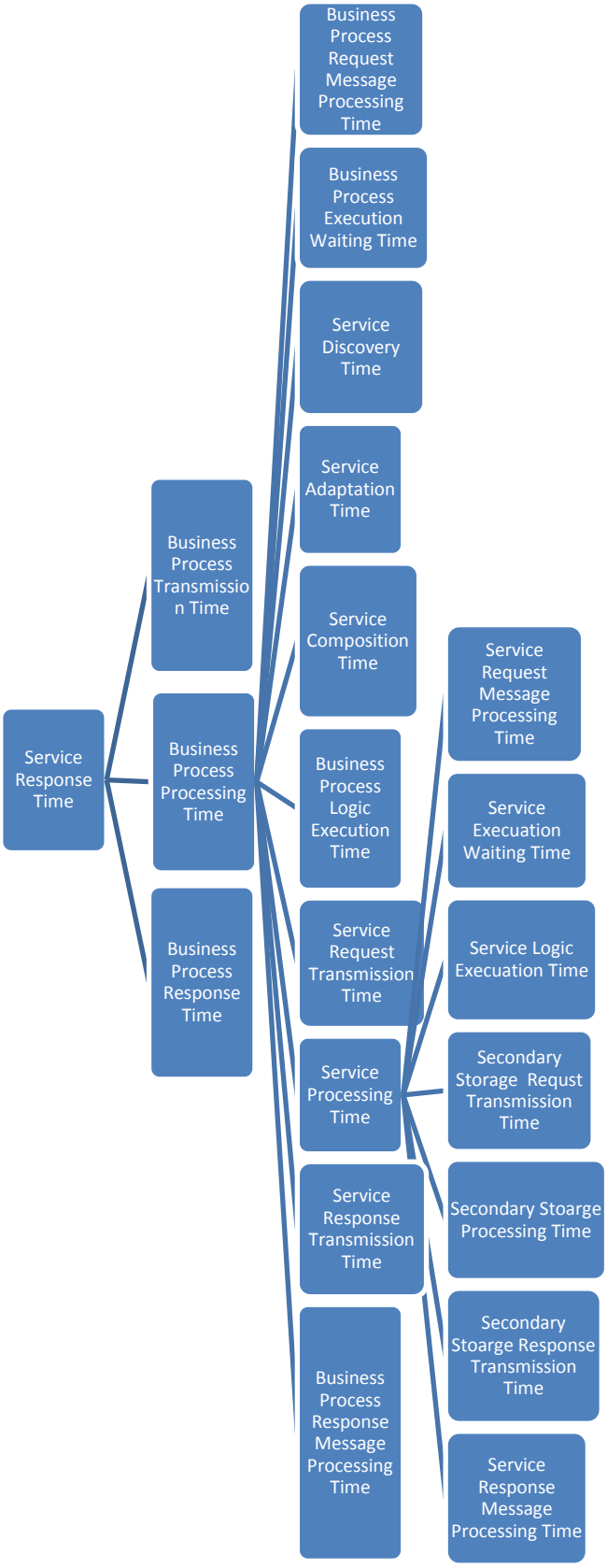


Figure. 2 Sub-metrics of SOA Response Times

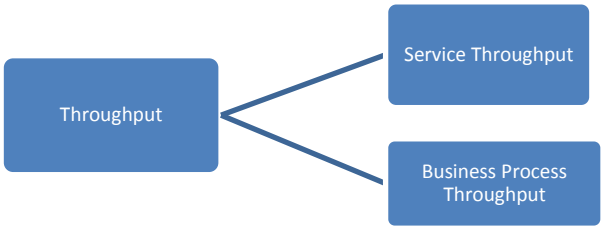


Figure. 3 Sub-metrics of SOA Throughput

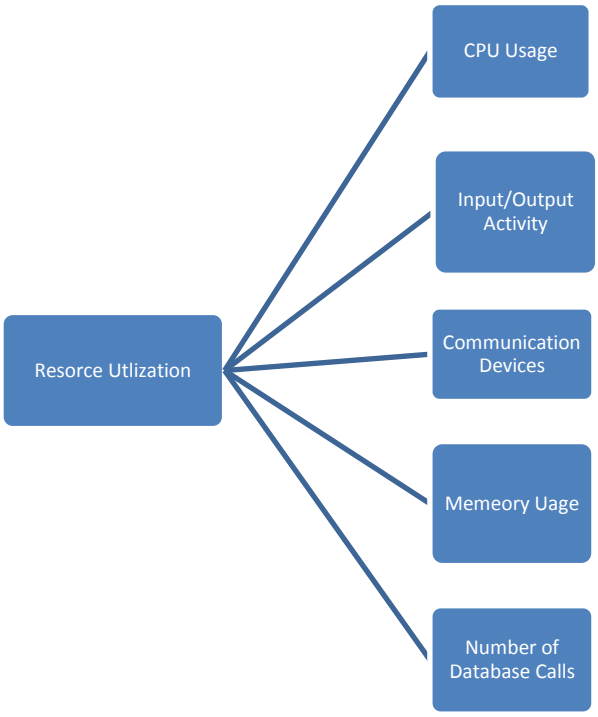


Figure. 4 Sub-metrics of SOA Resource Utilization

5. SOA PREDICTION AND EVALUATION APPROACHES

Several approaches have been created to evaluate and predict SOA based application performance. In the following we provide summaries of SOA performance prediction approaches in the scope of the survey. We have divided the approaches on seven columns such as author name and year of publication, main objective, prediction approach, analyzed metrics, performance model, validation method, and applicable project stage.

Table 1. Comparison of Several Prediction Approaches

Author Name/ Year	Main Objective	Approach used	Metrics Analyzed	Performance Model	Method's Validation	Applicable Project Stage
Kounev, Samuel, et al. [6], 2010	Designing systems with build-in self-aware performance and resource management capabilities	Use dynamic architecture-level performance model at run-time for online performance and resource management	Response time and Resource utilization	Queuing Petri net Model	Compared with PCM model result	Runtime
Liu, et al. [7], 2007	Develop a performance model for predicting runtime performance based on COTS ESB (Enterprise Service Bus)	Measure primitive performance overheads of service routing activities in the ESB	Throughput and response time	Queuing Network Model	Compared with the results of Microsoft Web Stress Tool	Runtime
Tribastone, et al. [8], 2010	Present a method for performance predication of SOA at early stage of development.	Modeling the system using UML and two profiles, UML4SOA, and MARTE	Response time, Processor utilization	Layered Queuing Network Model	Compared with Mobile Payment case study performance result	Design time
Teixeira, et al. [9], 2009	Propose approach to estimate performance of SOA	The model uses Petri Net formalism to represent the process and estimate its performance using simulation.	Resource consumption, Service levels degradation	Stochastic Petri Nets Model	Compared with (Rud et al) Analytical Method and values from real applications	Design time
Punitha, et al. [11], 2008	Developing an architectural performance model for SOA	Building and measuring the performance model using ACME language	Response time, throughput, load capacity, heavily loaded components.	Queuing Network Model	Prototype SOA Application has been implemented and measured	Design time
Brüseke, et al. [12], 2014	Developing PBlaman (Performance Blame Analysis)	Comparing the observed response time of each component in a failed test case to expected response time from the contract	Response time	Palladio Component Model (PCM)	Applied on two case studies	Design time
Reddy, et al. [13], 2011	Modeling Web Service using UML	Simulate the model using Simulation of Multi-tiered Queuing Applications (SMTQA)	Response time and Server utilization	SMTQA Model	Applied on case study	Design time

Marzolla, et al. [14], 2007	Present a multi-view approach for performance prediction of SOA based applications for users and providers	Approach for performance assessment of Web Service workflows described using annotated BPEL and WSDL specification	Response time and throughput	Queuing Network Model	Prototype tool called bpe12qnbound	Both Design time and Run time

6. CONCLUSION

We have surveyed the state-of-art in the research of performance prediction methods for service-oriented architecture based applications. The survey categorized the approaches according to the performance metrics analyzed, performance model, method validation, and approach applicable stage.

The field of performance evaluation and prediction for service-oriented architecture based application has been developed and matured over the last decade. Many tools and ideas have been implemented as good software engineering practice and should lead the creation of new approaches.

Our survey helps both architects and researchers. Architects can obtain a complete view of the performance evaluation and prediction approaches proposed to transfer them to industry, on the other hand researchers can align themselves with the proposed approaches and add more features in the future to enhance and enrich the area.

7. REFERENCES

[1] Bianco, P., Kotermanski, R., & Merson, P. F. (2007). Evaluating a service-oriented architecture.

[2] Krafzig, D., Banke, K., & Slama, D. (2005). Enterprise SOA: service-oriented architecture best practices. Prentice Hall Professional.

[3] Erl, T. (2004). Service-Oriented Architecture. Concepts, Technology, and Design. Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[4] Her, J. S., Choi, S. W., Oh, S. H., & Kim, S. D. (2007, October). A framework for measuring performance in service-oriented architecture. In Next Generation Web Services Practices, 2007. NWeSP 2007. Third International Conference on (pp. 55-60). IEEE.

[5] Abowd, G., Bass, L., Clements, P., Kazman, R., & Northrop, L. (1997).Recommended Best Industrial Practice for Software Architecture Evaluation (No. CMU/SEI-96-TR-025). CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST.

[6] Kounev, S., Brosig, F., Huber, N., & Reussner, R. (2010, July). Towards self-aware performance and resource management in modern service-oriented systems. In Services Computing (SCC), 2010 IEEE International Conference on(pp. 621-624). IEEE.

[7] Liu, Y., Gorton, I., & Zhu, L. (2007, July). Performance prediction of service-oriented applications based on an enterprise service bus. In Computer Software and Applications Conference, 2007. COMPSAC 2007. 31st Annual International(Vol. 1, pp. 327-334). IEEE.

[8] Tribastone, M., Mayer, P., & Wirsing, M. (2010). Performance prediction of service-oriented systems with layered queueing networks. In Leveraging Applications of Formal Methods, Verification, and Validation (pp. 51-65). Springer Berlin Heidelberg.

[9] Teixeira, M., Lima, R., Oliveira, C., & Maciel, P. (2009, October). Performance evaluation of service-oriented architecture through stochastic Petri nets. InSystems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on (pp. 2831-2836). IEEE.

[10] Balsamo, S., Mamprin, R., & Marzolla, M. (2004). Performance evaluation of software architectures with queuing network models. Proc. ESMc, 4.

[11] Punitha, S., & Babu, C. (2008, September). Performance prediction model for service oriented applications. In High Performance Computing and Communications, 2008. HPCC'08. 10th IEEE International Conference on (pp. 995-1000). IEEE.

[12] Brüseke, F., Wachsmuth, H., Engels, G., & Becker, S. (2014). PBlaman: performance blame analysis based on Palladio contracts. Concurrency and Computation: Practice and Experience.

[13] Reddy, C. R. M., Geetha, D. E., Srinivasa, K. G., Kumar, T. S., & Kanth, K. R. (2011). Predicting performance of web services using SMTQA. International Journal of Computer Science Information Technology, 1(2), 58-66.

[14] Marzolla, M., & Mirandola, R. (2007). Performance prediction of web service workflows. In *Software Architectures, Components, and Applications* (pp. 127-144). Springer Berlin Heidelberg.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 836 - 838, 2014, ISSN:- 2319–8656

Location Based Tracking System for Emergency Services

T. Swathi
Aurora's Technological and Reaserch Institute
Uppal, Hyderabad
India

B.S. Malleswari
Aurora's Technological and Research Institute
Uppal, Hyderabad
India.

Abstract Transmitting the geo-location information of a target via wireless networks is effective when both the target and the tracker are within Wi-Fi coverage area; the 802.11 wireless networks are not always accessible. When the target or the tracker is unable to access Wi-Fi, it is impossible to perform location tracking. Therefore, SMS is a relatively more reliable and flexible solution because of its widespread use. In this system, a device is equipped with a global system for mobile communications (GSM) modem and a GPS unit. It transmits short messages containing its GPS coordinates to the server at 30-s intervals. Although transmitting the geo-location information of a target via wireless networks is effective when both the target and the tracker are within Wi-Fi coverage area, the 802.11 wireless networks are not always accessible. When the target or the tracker is unable to access Wi-Fi, it is impossible to perform location tracking. In this System, a novel method called location-based delivery (LBD), which combines the short message service (SMS) and global position system (GPS). LBD reduces the number of short message transmissions while maintaining the location tracking accuracy within the acceptable range. The proposed approach, LBD, consists of three primary features: Short message format, location prediction, and dynamic threshold. The defined short message format is proprietary.

Key Words: Short Message Service (SMS), Location Tracking, Mobile Phones, Prediction Algorithms, Global Positioning System (GPS).

1. INTRODUCTION

Location based tracking and handling the devices is based on the global position system (GPS) is common in the growing world, and therefore, several location tracking applications have been developed, including continuous location based transport, system or vehicle based intelligent transport, monitoring vehicles, tracking elders, children's and women employees for their safety reasons or to prevent them from the being lost. The GPS is mainly used to obtain geographical location of the object (e.g., a transmitter devices or mobile devices). However, most of the above-citedworks used either an 802.11 wireless network or the short messageservice (SMS) to transmit the location information of a targetto a tracker. Real time tracking system is majorly used for care management applications for

children and mentally challenged people; the main aim of the system is to transfer the location and position of the objective to the mobile device to a central GPS application server through the 802.11 wireless networks. This application allows the server to simultaneously monitor multiple targets (e.g., elders or children), this is in line with Lee et al.Further, Choi et al. assumed that the location information of a targets transmitted through wireless networks. Their work focused on proposing a geolocation update scheme to decrease the update. Frequency. Lita et al. proposed an automobile localization system by using SMS.The proposed system, which is interconnectedwith the car alarm system, transmits alerts to theowner's mobile phone in the event of a car theft (e.g., activationof the car alarm, starting of the engine) or provides informationfor monitoring adolescent drivers (e.g., exceeding the speedlimit or leaving a specific area). Hameed et al. proposed a carmonitoring and tracking system that uses both SMS and GPSto prevent car theft. Anderson et al. proposed a transportationinformation system . In this system, a hardware device called Star Box, which is equipped with a global system for mobile communications (GSM) modem and a GPS unit, is installed in a vehicle to track the vehicle's location. Star Box transmits short messages containing its GPS coordinates to the server at30-s intervals. The users can send short messages to the server to determine the expected arrival time of buses at their locations. Although transmitting the geolocation information of a targetvia wireless networks is effective when both the target and the tracker are within Wi-Fi coverage area, the 802.11 wireless networksare not always accessible. When the target or the trackeris unable to access Wi-Fi, it is impossible to perform locationtracking. Therefore, SMS is a relatively more reliable and flexiblesolution because of its widespread use (i.e., well-structuredworldwide) [6], [8]. However, SMS is a user-pay service.

The objective of this study is to minimize the transmission cost ofa tracking system by minimizing the number of SMS transmissionswhile maintaining the location tracking accuracy.

In this paper, a novel method called location-based delivery (LBD), which combines SMS and GPS, is proposed, and further, a realistic system to perform precise location trackingis developed. LBD mainly applies the following two proposedtechniques: Location prediction and dynamic threshold. Locationprediction is performed by using the current location, movingspeed, and bearing of the target to predict its next location.When the distance between the predicted location and theactual location exceeds a certain

threshold, the target transmitsa shortmessage to the tracker to update its current location. Thedynamic threshold maintains the location tracking accuracy andnumber of short messages on the basis of the moving speed ofthe target.The simulations performed to test the performance of LBD show that compared with other related works; the proposed LBD minimizes the number of short message transmissions while maintaining the location prediction accuracy within the acceptable range.

Figure-1 Overview of location based tacking

Figure-1 Overview of location based tracking.

2. SHORT MESSAGE SERVICE

SMS is a text messaging service component of phone, web, or mobile communication systems, using standardized communications protocols that allow the exchange of short text messages between fixed line or mobile phone devices. SMS text messaging is the most widely used data application in the world, with 3.6 billion active users, or 78% of all mobile phone subscribers. A short message is transmitted from the mobile station (MS) to the GSM base station (BTS) through a wireless link and is received in the backbone network of the service provider. The mobile switch center (MSC), home location register (HLR), and visitor location register (VLR) determine the appropriate short message service center (SMSC), which processes the message by applying the "store and forward" mechanism.

The term SMS is used as a synonym for all types of short text messaging as well as the user activity itself in many parts of the world. SMS is also being used as a form of direct marketing known as SMS marketing. SMS as used on modern handsets originated from radio telegraphy in radio memo pagers using standardized phone protocols and later defined as part of the GSM (Global System for Mobile Positioning) series of standards in 1985 as a means of sending messages of up to 160 characters, to and from GSM mobile handsets. Since then, support for the service has expanded to include other mobile technologies such as ANSI CDMA networks and digital AMPs, as well as satellite and landline networks. Most SMS messages are mobile-to-mobile text messages though the standard supports other types of broadcast messaging as well.

3. LOCATION BASED DELIVERY

Location based services (LBS) a novel class of computer application, which combines the short message services (SMS) and global position system (GPS). As such LBS is an informative service and number of uses in social networking today as an entertainment service, which is accessible with mobile devices through the mobile network and which uses information on the geographical position of the mobile device.

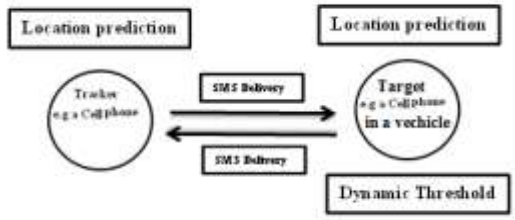


Figure-2 Structure of the LBD systeem.

The proposed approach, LBD, consists of three primary features: Shortmessage format, location prediction, and dynamic threshold. The defined short message format is proprietary. Location prediction isperformed by using the current location, moving speed, and bearing of the target to predict its next location. When the distance between the predicted location and the actual location exceeds a certain threshold, the target transmits a short message to the trackerto update its current location. The threshold is dynamically adjusted to maintain the location tracking accuracy and the numberof short messages on the basis of themoving speed of the target. It satisfactorily maintains the location tracking accuracy with relatively fewer messages.The threshold is dynamically adjusted to maintain the location tracking accuracy.

4. CONCLUSION:

In this System, a novel method called location-based delivery (LBD), which combines the short message service (SMS) and global position system (GPS). LBD reduces the number of short message transmissions while maintaining the location tracking accuracy within the acceptable range. The proposed approach, LBD, consists of three primary features: Short message format, location prediction, and dynamic threshold. The defined short message format is proprietary. Location prediction is performed by using the current location, moving speed, and bearing of the target to predict its next location. When the distance between the predicted location and the actual location exceeds a certain threshold, the target transmits a short message to the trackerto update its current location. The threshold is dynamically adjusted to maintain the location tracking accuracy and the number of short messages on the basis of the moving speed of the target.

5. REFERENCES

[1] H. H. Lee, I. K. Park, and K. S. Hong, "Design and implementation ofa mobile devices-based real-time location tracking," in *Proc. UBICOMM*, 2008, pp. 178–183.

[2] Z. Tian, J. Yang, and J. Zhang, "Location-based services applied to aelectric wheelchair based on the GPS and GSM networks," in *Proc. ISA*,2009, pp. 1–4.

[3] I. Lita, I. B. Cioc, and D. A. Visan, "A new approach of automobile localizationsystem using GPS and GSM/GPRS transmission," in *Proc. ISSE*,2006, pp. 115–119.

[4] P. Perugu, "An innovative method using GPS tracking, WINS technologiesfor border security and tracking of vehicles," in *Proc. RSTSCC*, 2010, pp. 130–133.

[5] S. A. Hameed, O. Khalifa, M. Ershad, F. Zahudi, B. Sheyaa, and W.Asender, "Car monitoring, alerting, and tracking model: Enhancementwith mobility and database facilities," in *Proc. ICCCE*, 2010, pp. 1–5.

[6] R. E. Anderson, A. Poon, C. Lustig, W. Brunette, G. Borriello, and B. E.Kolko, "Building a transportation information system using only GPS andbasic SMS infrastructure," in *Proc. ICTD*, 2009, pp. 233–242.

[7] W. J. Choi and S. Tekinay, "Location-based services for next-generationwireless mobile networks," in *Proc. IEEE VTC*, 2003, pp. 1988–1992.

[8] R. E. Anderson, W. Brunette, E. Johnson, C. Lustig, A. Poon, C. Putnam,
O. Salihbaeva, B. E. Kolko, and G. Borrielllo, "Experiences with atransportation information system that uses only GPS and SMS," in *Proc.ICTD*, 2010.

[9] A. Civilis, C. S. Jensen, and S. Pakalnis, "Techniques for efficient roadnetwork-based tracking of moving objects," *IEEE Trans. Knowl.DataEng.*, vol. 17, no. 5, pp. 698–712, 2005.

[10] M. Zahaby, P. Gaonjur, and S. Farajian, "Location tracking in GPS usingKalman filter through SMS," in *Proc. IEEE EUROCON*, 2009, pp. 1707–1711.

[11] A. Civilis, C. S. Jensen, J. Nenortaite, and S. Pakalnis, "Efficient trackingof moving objects with precision guarantees," in *Proc. MOBIQUITOUS*,2004, pp. 164–173.

[12] Y. Y. Xiao, H. Zhang, and H. Y. Wang, "Location prediction for trackingmoving objects based on grey theory," in *Proc. FSKD*, 2007, pp. 390–394.

[13] P. H. Tseng, K. T. Feng, Y. C. Lin, and C. L. Chen, "Wireless locationtracking algorithms for environments with insufficient signal sources,"*IEEE Trans. Mobile Comput.*, vol. 8, no. 12, pp. 1676–1689, 2009.

[14] R. Bajaj, S. L. Ranaweera, and D. P. Agrawal, "GPS: Location-trackingtechnology," *Computer*, vol. 35, no. 4, pp. 92–94, 2002.

[15] Movable Type Scripts. (2012 June).[Online]. Available: http://www.mova
ble-type.co.uk/scripts/latlong.html

International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 839 - 843, 2014, ISSN:- 2319–8656

Spam filtering by using Genetic based Feature Selection

Sorayya Mirzapour Kalaibar
Department of Computer, Shabestar Branch
Islamic Azad University
Shabestar, Iran

Seyed Naser Razavi
Computer Engineering Department
Faulty of Electrical and Computer Engineering,
University of Tabriz, Iran

Abstract: Spam is defined as redundant and unwanted electronical letters, and nowadays, it has created many problems in business life such as occupying networks bandwidth and the space of user's mailbox. Due to these problems, much research has been carried out in this regard by using classification technique. The resent research show that feature selection can have positive effect on the efficiency of machine learning algorithm. Most algorithms try to present a data model depending on certain detection of small set of features. Unrelated features in the process of making model result in weak estimation and more computations. In this research it has been tried to evaluate spam detection in legal electronica letters, and their effect on several Machin learning algorithms through presenting a feature selection method based on genetic algorithm. Bayesian network and KNN classifiers have been taken into account in classification phase and spam base dataset is used.

Keywords: Email spam, feature selection, genetic algorithm, classification.

1. INTRODUCTION

Nowadays, e-mail has been widely considered as one of the fastest and most economical forms of communication .Thus, the e-mail is prone to be misused. such misuse is posting unsolicited, unwanted e-mails known as spam or junk e-mails[1]. Spam has been considered as a serious problem. Many Internet Service Providers (ISPs) receive more than billion spam messages per day. Much of these e-mails are filtered before end users can access them. Content-Based filtering is a key technological method for e-mail filtering. The spam e-mail contents usually contain common words called features. Frequency of occurrence of these features inside an e-mail gives an indication that the e-mail is a spam or legitimate [2,3,4]. There are various purposes in sending spams such as economical purposes. Some spams are unwanted advertising and commercial messages, while others deceive the users to use their private information (phishing), or they temporarily destroy the mail server by sending malicious software to the user's computer. Also, they create traffic, or distribute immoral messages. Therefore, it is necessary to find some ways to filter these troublesome and annoying emails automatically. In order to detect spams, some methods such as parameter optimization and feature selection methods have been proposed in order to reduce processing overhead and to guarantee high detection rate [16].The spam filtering is high sensitive application of text classification (TC) task. The main problem in text classification tasks which is more serious in email filtering is existence of large number of features. For solving the issue, various feature selection methods are considered. They extract one, and offer it as input to classifier[5]. In this paper, we incorporate genetic algorithm to find an optimal subset of features of the spam base data set. The selected features are used for classification of the spam base.

2. LITERATURE REVIEW

Features selection approaches are usually employed to reduce the size of feature set, and to select a subset of original features. Over the past years, the following methods have been considered to select effective features such as the algorithms based on population to select important features, and to remove irrelevant and redundant features such as genetic algorithm (GA), particle swarm optimization (PSO), and ant colony algorithm (ACO). Some algorithms are developed to classify and filter e-mails. The RIPPER algorithm [6] is a rule-based algorithm used for filtering e-mails. Drucker, et. al. [7] proposed an SVM algorithm for spam categorization. Sahami, et. al. [8] proposed Bayesian junk E-mail filter using bag-of-words representation and Naïve Bayes algorithm. Clark, et. al. [9] used the bag-of-words representation and ANN for automated spam filtering system. Branke, J. [10] discussed how genetic algorithm can be used to assist designing and training. Riley. J. [11] described a method of utilizing genetic algorithms to train fixed architecture feed-forward and recurrent neural networks. Yao. X. and Liu. Y. [12] reviewed different combinations between ANN and GA, and used GA to evolve ANN connection weights, architectures, learning rules, and input features. Wang and et al. presented feature selection incorporation based on genetic algorithm and support vector machine based on SRM to detect spam and legitimate emails. The presented method had better results than main SVM [13]. Zhu developed a new method based on rough set and SVM in order to improve the level of classification. Rough set was used as a feature selection to decrease the number of feature and SVM as a classifier [14]. Fagboula and et al. considered GA to select an appropriate subset of features, and they used SVM as a classifier. In order to improve the classification accuracy and computation time, some experiments were carried out in terms of data set of Spam assassin [15]. Patwadhan and Ozarkar presented random forest algorithm and partial decision trees for spam classification. Some feature selection methods have been used as a preprocessing stage such as Correlation based feature selection, Chi-square, Entropy, Information Gain, Gain Ratio, Mutual Information, Symmetrical Uncertainty, One R and Relief. Using above mentioned methods resulting in selecting more efficient and useful features decrease time complexity and increase accuracy [17].

3. GENETIC ALGORITHMS

A genetic algorithm (GA) is one heuristic techniques that are based on natural selection from the population members, and tries to find high-quality solutions to large and complex optimization problems. This algorithm can identify and exploit regularities in the environment, and converges on solutions (it can also be regarded as locating the local maxima) that were globally optimal [18]. This method is very effective and widely used to find-out optimal or near optimal solutions to a wide variety of problems. The genetic algorithm repeatedly modifies the population of individual solutions. At each step, the genetic algorithm tries to select the best individuals. Now, "parent" population genetic algorithm creates "children" constituting next generation. Over successive generations, the population evolves toward an optimal solution. The genetic algorithm uses three main rules at each step to create next generation: a. Select the individuals, called parents that contribute to the population at the next generation. b. Crossover rules that combine two parents to form children for the next generation. c. Mutation rules, apply random changes to individual parents to form children

4. FEATURE SELECTION

Features selection approaches are usually employed to reduce the size of feature set, and to select a subset of the original features. We use the proposed genetic algorithms to optimize the features that significantly contribute to the classification.

4.1. Feature Selection Using Proposed Genetic Algorithm

In this section, the method of feature selection by using the proposed genetic Algorithm has been presented. The procedure of the proposed method has been stated in details in the following section.

4.1.1. Initialize population

In the genetic algorithm, each solution to the feature selection problem is a string of binary numbers called chromosome. In this algorithm, initial population is generated randomly. IN feature representation is considered as a chromosome, and if the value of chromosome [i] is 1, the ith feature is selected for classification, while if it is 0, then these features will be removed [19,20]. Figure 1 shows feature presentation as a chromosome.

Chromosome:

F_1	F_2	F_3	...	F_{n-1}	F_n
1	0	1	...	1	0

Figure 1. Feature Subset: $\{F_1, F_3, \ldots, F_{n-1}\}$

In this research, we used weighted F-score to calculate the fitness value of each chromosome. The algorithm starts by randomly initializing a population of N number of initial chromosome.

4.1.2 Cross over

The crossover is the most important operation in GA. Crossover, as its name suggests, is a process of recombination of bit strings via the exchange of segments between pairs of chromosomes. There are various kinds of crossover. In one point of cross-over, a bit position is randomly selected that should be changed. In this process, a random number is generated. This number (less than or equal to the chromosome length) is the crossover position [21]. Here, one crossover point is selected, binary string from beginning of chromosome to the crossover point is copied from one parent, and the rest is copied from the second parent [22].

4.1.3. Proposed mutation

In mutation, it can be ensured that all possible chromosomes can maintain good gene in the newly generated chromosomes. In our approach, Mutation operator is a two-steps process, and is a combination of random and substitution mutation operator. Also it occurs on the basis of two various mutation rates. In mutation operator, substitution step is considered with the probability of 0.03. In each generation, the best chromosome involving better features and higher fitness is selected, and it substitutes for the weakest chromosome having lesser fitness than others. In this stage, the better chromosome transfers the current generation to next generation, and it follows rapid convergence of algorithm. Otherwise, it enters the second mutation step with probability of 0.02. This step changes some gens of chromosome randomly by inverting their binary cells. In fact, the second one is considered to prevent reducing exploration capability of search space to keep diversity in other chromosomes. Generally, mutation probability is equal to 0.05.

5. RESULTS SIMULATION

In order to investigate the impact of our approach on email spam classification, spam base data set downloaded from the UCI Machine Learning Repository has been used [23]. Data set of Spam base involving 4601 emails was proposed by Mark Hopkinsand his colleagues. In This data set that is divided into two parts, 1 shows spam, and zero indicates non-spam. This data set involves 57 features with continuous values. In simulation of the proposed method, training set involving 70% of the main data set and two experimental sets have been separately considered for feature selection and classification. Each one involves 15% of the main data set. After performing feature selection by using the training set, the test set was used to evaluate the selected subset of features. The evaluation of overall process was based

on weighted f-score which is a suitable measure for the spam classification problem. The performance of spam filtering techniques is determined by two well-known measures used in text classification. These measures are precision and recall [24, 25]. Here four metric have been used for evaluating the performance of proposed method such as precision, accuracy, recall and F1 score. These metrics are computed as follows:

$$\pi_i = \frac{TP_i}{TP_i + FP_i} \quad (1)$$

$$\rho_i = \frac{TP_i}{TP_i + FN_i} \quad (2)$$

$$F1 = \frac{2 \times \pi \times \rho}{(\pi + \rho)} \quad (3)$$

$$Accuracy = \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i} \quad (4)$$

Where:
TP_i = the number of test samples that have been properly classified in c_i class.
FP_i = the number of test samples that have been incorrectly classified in c_i class.
TN_i = the number of test samples belonging to c_i class, and have been correctly classified in other classes.
FN_i = the number of test samples belonging to c_i class, and have been incorrectly classified in other classes.
The methods of Bayesian network and K nearest neighbors algorithm (KNN) have been used for classification. The executed program and the obtained average have been compared 8 times to investigate the performance of each classifier. The results obtained from the proposed method of feature selection have been compared without considering feature selection. The obtained results show that when the parameters are presented in tables 1, the best performance is observed in terms of GAFS.

Table 1: the parameters of feature selection by using genetic algorithm

Initial population	80
Mutation rate1	0.03
Mutation rate2	0.02
Crossover	0.7
Generations	100

6. RESULT EVALUATION

In this section, the results of experiments have been presented to evaluate efficiency of the proposed method. The results of comparing classifier Bayesian network and KNN have been presented in table 2. In addition, figure 2 shows graphical diagram of the effects of the proposed method on reduction of redundant features. According to the results obtained in terms of Bayesian network classification, the proposed method has increased the classification accuracy, and at the same time, it has removed significant number of features. Also, although KNN classifier has removed some features, it has reached the accuracy that is near to the accuracy obtained before selecting the feature. The results obtained for three other criteria have been demonstrated in table 3. As it can be observed in table, Bayesian classifier has been considerably optimized in all three evaluation criteria, while KNN classifier has reached to the precision near to the previous precision. These results indicate that feature selection by GA technique improves email spam classification. GA FS and all features by using mentioned classifiers have been compared in terms of accuracy, number of selected feature, recall, precision and F score of spam class

Table 2: comparing feature selection methods in terms of accuracy

Algorithms / classifier	All Feature	GA FS
Bayesian network	0.891	0.918
KNN (N=1)	0.9	0.891

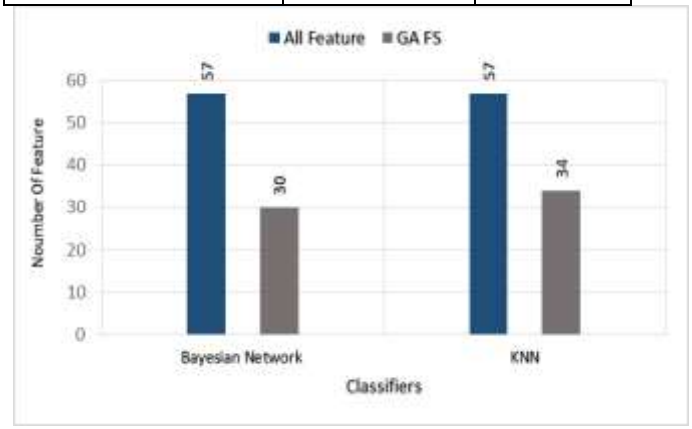


Figure 3: column graph of comparing the number of selected features

Table 3: comparing feature selection methods

classifier / measures	KNN(N=1)		Bayesian network	
	All Feature	**GA FS**	**All Feature**	**GA FS**
precision	0.892	0.886	0.89	0.935
recall	0.871	0.860	0.851	0.869
F1 score	0.882	0.871	0.87	0.900

7. CONCLUTION

In this paper, the proposed GA based feature selection method has been presented and evaluated by using data set of Spam Base. The results obtained from proposed method were compared with position without feature selection. The obtained results show that, with regard to the number of removed features, the proposed method has accuracy comparable with the methods that lack feature selection. In addition, in Bayesian network classifier, better results have been obtained compared to KNN classifier and all evaluation criteria have been considerably improved. Therefore, the proposed method has considerable effect on features selection and increasing the accuracy. We can use parameter optimization in this work. Also, the proposed algorithm can be combined with other classification algorithms in the future.

REFERENCE

[1] GOWEDER, A. M., RASHED, T., ELBEKAIE, A., & ALHAMMI, H. A. (2008). An Anti-Spam System Using Artificial Neural Networks and Genetic Algorithms. Paper presented at the Proceedings of the 2008 International Arab Conference on Information Technology.

[2] Bruening, P.(2004). Technological Responses to the Problem of Spam: Preserving Free Speech and Open Internet Values. First Conference on E-mail and Anti-Spam.

[3] Graham, P.(2003). A Plan for Spam. MIT Conference on Spam.

[4] William, S., et. al. (2005). A Unified Model of Spam Filtration, MIT Spam Conference, Cambridge.

[5] GOWEDER, A. M., RASHED, T., ELBEKAIE, A., & ALHAMMI, H. A. (2008). An Anti-Spam System Using Artificial Neural Networks and Genetic Algorithms. Paper presented at the Proceedings of the 2008 International Arab Conference on Information Technology.

[6] Cohen, W. (1996). Learning Rules that Classify E-mail, In AAAI Spring Symposium on Machine Learning in Information Access, California.

[7] Drucker, H., et. al.(1999) Support Vector Machines for Spam Categorization, In IEEE Transactions on Neural Networks.

Sahami, M., et. al.,(1998). A Bayesian Approach to Filtering Junk E-Mail, In Learning for Text Categorization, AAAI Technical Report, U.S.A.

[8] Riley. J. (2002). An evolutionary approach to training Feed-Forward and Recurrent Neural Networks", Master thesis of Applied Science in Information Technology, Department of Computer Science, Royal Melbourne Institute of Technology, Australia.

[9] Clark, et. al. (2003). A Neural Network Based Approach to Automated E-Mail Classification, IEEE/WIC International Conference on Web Intelligence.

[10] Branke, J. (1995). Evolutionary algorithms for neural network design and training, In Proceedings 1st Nordic Workshop on Genetic Algorithms and its Applications, Finland.

[11] Yao. X., Liu. Y. (1997). A new evolutionary system for evolving artificial neural networks", IEEE Transactions on Neural Networks.

[12] Wang, H.-b., Y. Yu, and Z. Liu. (2005) SVM classifier incorporating feature selection using GA for spam detection, in Embedded and Ubiquitous Computing–EUC 2005., Springer. p. 1147-1154.

[13] Zhu, Z. (2008). An email classification model based on rough set and support vector machine. in Fuzzy Systems and Knowledge Discovery.

[14] .Temitayo, F., O. Stephen, and A. Abimbola. (2012). Hybrid GA-SVM for efficient feature selection in e-mail classification. Computer Engineering and Intelligent Systems. 3(3): p. 17-28.

[15] Stern, H. (2008) A Survey of Modern Spam Tools. in CEAS. Citeseer.

[16] Ozarkar, P. and M. Patwardhan. (2013). INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET). Journal Impact Factor. 4(3): p. 123-139.

[17] Zhang, L., Zhu, J., & Yao, T. (2004). An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing (TALIP), 3(4), 243-269.

[18] Vafaie H, De Jong K. (1992). Genetic algorithms as a tool for feature selection in machine learning. In Proceedings of Fourth International Conference on Tools with Artificial Intelligence (TAI '92). 200-203.

[19] Yang J, Honavar V. (1998). Feature subset selection using a genetic algorithm. Intelligent Systems and their Applications, IEEE, 13(2):44-49.

[20] Shrivastava, J. N., & Bindu, M. H. (2014). E-mail Spam Filtering Using Adaptive Genetic Algorithm. International Journal of Intelligent Systems & Applications, 6(2).

[21] Karimpour, J., A.A. Noroozi, and A. Abadi. (2012). The Impact of Feature Selection on Web Spam Detection. International Journal of Intelligent Systems and Applications (IJISA), 4(9): p. 61.

[22] UCI repository of Machine learning Databases. (1998). Department of Information and Computer Science, University of California, Irvine, CA, http://www.ics.uci.edu/~mlearn/MLRepository.html, Hettich, S., Blake, C. L., and Merz, C. J.

[23] Liao, C., Alpha, S., Dixon.P. (2004). Feature Preparation in Text Categorization, Oracle Corporation.

[24] Clark, et. al. (2003). A Neural Network Based Approach to Automated E-Mail Classification, IEEE/WIC International Conference on Web Intelligence