International Journal of Computer Applications Technology and Research
Volume 3– Issue 3, 136 - 139, 2014, ISSN: 2319–8656

Image Steganography Using HBC and RDH Technique

Hemalatha .M
Sri Manakula Vinayagar Engineering College
Pudhucherry, India

Prasanna.A
Sri Manakula Vinayagar Engineering College
Pudhucherry, India

Dinesh Kumar R
Sri Manakula Vinayagar Engineering College
Pudhucherry, India

Vinoth kumar D
Sri Manakula Vinayagar Engineering College
Pudhucherry, India

Abstract**:** There are algorithms in existence for hiding data within an image. The proposed scheme treats the image as a whole. Here Integer Cosine Transform (ICT) and Integer Wavelet Transform (IWT) is combined for converting signal to frequency. Hide Behind Corner (HBC) algorithm is used to place a key at corners of the image. All the corner keys are encrypted by generating Pseudo Random Numbers. The Secret keys are used for corner parts. Then the hidden image is transmitted. The receiver should be aware of the keys that are used at the corners while encrypting the image. Reverse Data Hiding (RDH) is used to get the original image and it proceeds once when all the corners are unlocked with proper secret keys. With these methods the performance of the stegnographic technique is improved in terms of PSNR value.

Keywords: ICT, IWT, HBC, RDH, Pseudo Random Number, Secret Key.

1. INTRODUCTION

One of the successful reasons behind the intruders to acquire the data easily is due to the reason that the system is in a form that they can read and comprehend the data. Intruders may reveal the information to others, modify it to misrepresent an individual or organization, or use it to launch an attack. One solution to this problem is, through the use of steganography. Steganography is a technique of hiding information in digital media. In contrast to cryptography, it is not to keep others from knowing the hidden information but it is to keep others from thinking that the information even exists. Steganography become more important as more people join the cyberspace revolution. Due to advances in ICT, most of information is kept electronically. The host data set is purposely corrupted, but in a covert way, designed to be invisible to an information analysis.

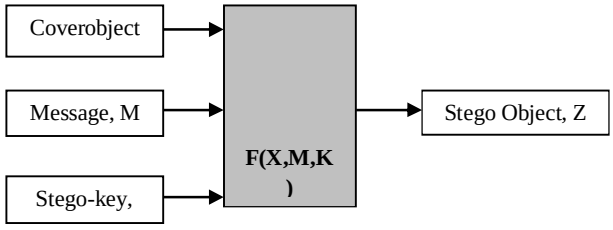

Figure 1: Encryption of an Image

There are many methods that can be used to detect Steganography such as: "Viewing the file and comparing it to another copy of the file found on the Internet (Picture file). The Proposed System consists of the different methods to be used in the encryption and the *data hiding* and *the retrieval phase*. The data hiding phase consist of the RDH method which is used to hide the data in different format and can be extracted using the different technique. The *Region separation method* is used to hide the secret data in the different region of the image and so ,only the authorized user can decrypt and access the data. The ICT and IWT methods are used to hide the data in the image so that the original image is not altered. The mechanism used to protect the loss of data by cropping the stegno image that contains the data is RDH so that image cannot be cropped. The security level for the data is increased in this kind of system.

2. RELATED WORKS

On the part of stegnography 'n' number of works has been developed. In the encryption phase the data carrying pixel should be hidden. Our proposed work provide these to increase the secrecy of the data. Katzenbeisser, S. and Petitcolas, F.A.P., [1] proposed Information Hiding Techniques for Steganography and Digital Watermarking. It helps in copyright protection. M. F. Tolba, M. A. Ghonemy, I. A. Taha, A. S. Khalifa [2] proposed Integer Wavelet Transforms in Colored Image-Stegnography. The frequency and the location information is captured. Guorong Xuan et. al [3] proposed Distortionless Data Hiding Based on Integer Wavelet Transform. It provides. Shejul, A. A., Kulkarni, U.L.,[4] proposed A Secure Skin Tone based Steganography (SSTS) using Wavelet Transform. cropping case used here preserves histogram of DWT coefficients after embedding. It can be used aldo to prevents histogram based attacks. Masud, Karim S.M., Rahman, M.S., Hossain, M.I. [5] proposed A New Approach for LSB Based Image Steganography using Secret Key. It is difficult to extract the hidden information knowing the retrieval methods. The Peak Signal-to-Noise Ratio (PSNR) measures the quality of the stego images and also gives better result. This is because of very small number of bits of the image.

Xie, Qing., Xie, Jianquan., Xiao, Yunhua. [6] A High Capacity Information Hiding Algorithm in Color Image. The security is much higher because the visual effect of image is not affected. Sachdeva, S and Kumar, A., [7] Colour Image Steganography Based on Modified Quantization Table. The cover image is divided into blocks and DCT is applied to each block. IDCT is applied to produce the stego image which is identical to cover image. Chen, R. J., Peng, Y. C., Lin, J. J., Lai, J. L., Horng, S. J. [8] Multi-bit Bitwise Adaptive Embedding Algorithms with Minimum Error for Data Hiding. The system provides embedding algorithms that results in minimum error and it is suitable to hardware implementation due to it is based on logic, algebraic, and bit operations. Roy, S., Parekh, R., [9] A Secure Keyless Image Steganography Approach for Lossless RGB Images. The system authentication is provided and Storage capacity is increased. Hiding the information provides minimal Image degradation.. Mandal, J.K., Sengupta, M., [10] Steganographic Technique Based on Minimum Deviation of Fidelity (STMDF). It shows better performance in terms of PSNR and fidelity of the stego images.

3. SYSTEM ARCHITECTURE

The system architecture or the design gives value of revealing the process that is done during the experimental works. The sender first authenticates himself to enter the system which is known as the login details that is stored in the database and the takes the image that he wants to transmit and collects the data that are important as a cover message and then encrypts the image. A key is provided. This stegno image will be transmitted over the networks and it will be recovered in the receiver end. Then the original secret data is said to be constructed and then the original image and hidden data can be regained by using the absolute keys.

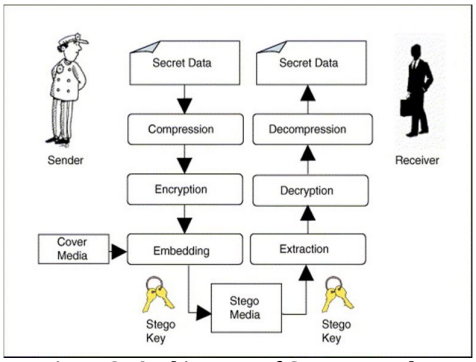

Figure 2: Architecture of Steganography

4. RESEARCH PROPOSAL

STEP 1: CLASSIFYING INTO PIXELS

Here ICT and IWT are used to split the image into pixels. A Integer cosine transform (ICT) expresses a finite sequence of data points in terms of a sum of cosine functions oscillating at different frequencies. An Integer wavelet transform (IWT) is any wavelet transform for which the wavelets are discretely sampled. Temporal resolution is maintained. The pixels are initially classified and then data for each of the pixel is embedded. This increases the confidentiality of the data that is to be hidden and transmitted.

Algorithm 1: ICT

The integer cosine transform (ICT) is an approximation of the discrete cosine transform.Integer arithmetic mode is used in implementation. It promotes the cost and speed of hardware implementation.

```
if (temp == 255)
{
  i++;
  int value = ICT[i],  length = ICT[i];
 for(j=0; j<length; j++)
 {
   pixel[k] = value;
   k++;
```

Algorithm 2: IWT

This algorithm is used to reduce the space of usage. This part is also associated with classifying the pixels of an image. The area without a pixel value or RGB value is skipped.

```
IWT( )
while( h >= minWaveLength )
 {
double[ ] iBuf = new double[ h ];
 for( int i = 0; i < h; i++ )
{
 iBuf[ i ] = arrHilb[ i ];
double[ ] oBuf = _wavelet.forward( iBuf );
}
 for( int i = 0; i < h; i++ )
{
arrHilb[ i ] = oBuf[ i ];
 h = h >> 1;
 level++;
```

STEP 2: GENERATING RANDOM NUMBERS AT THE CORNERS

Here a new least significant bit embedding algorithm for hiding secret messages in non adjacent pixel locations of edges in the image is proposed. Here the messages are hidden in regions which are least like their neighboring pixels so that an attacker will have less suspicion of the presence of message bits in edges, because pixels in edges appear to be much brighter or dimmer than their neighbours. Edges can be detected by edge detection filters . For a 3x3 window Laplacian edge detector has the following form.

$$D=8x_5 - (x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9)$$

Where x_1 , x_2 , x_3 , x_4 , x_5 , x_6, x_7, x_8, x_9 and are the pixel values in a sliding 3x3 window scanning from the top left to bottom right with center pixel value.

x= D will become positive when the center pixel x is brighter is brighter than its neighbours and vice versa. The disadvantage of LSB embedding is that it creates an imbalance between the neighbouring pixels causing the value of D to change. Here this imbalance is avoided by flipping the gray-scale values among 2i-1, 2i and 2i+1. D after LSB embedding is not different from the old value of D before embedding. The various strengths of this scheme are that an

attacker will have less suspicion to the presence of message bits in edges because pixels in edges appear to be either much brighter or dimmer than their neighbours and it is also secure against blind steganalysis. In order to ensure that the neighbouring pixels in the window are not changed by Laplacian edge detectors, we apply the edge detection filter in non overlapping window only. It also limits the length of the secret message to be embedded. The proposed algorithm random edge LSB (RELSB) embedding uses least significant bit embedding at random locations in nonadjacent edge pixels of the image.

Algorithm 3: LSB

```
hideMessage()
{
string message = messageTextField.getText();
boolean displayInWhite = checkBox.getState();
if(originalImage == null)
{
Frame f = new Frame();
MessageDialog notL = new MessageDialog(f, "Error",
"Please load an image to hide the message");
notL.pack();
notL.show();
return;
}
if (message.length() == 0 || message.length()>40)
{
Frame f = new Frame();
MessageDialog mdialog = new MessageDialog(f, "Error",
"Please use a valid message (less than 40 letters)");
mdialog.pack();
mdialog.show();
return;
}
}
```

Figure 3: HBC Technique

A technique called *Pseudo Random Generation* is used here for generating numbers at the corners of image**.** It is for generating a sequence of numbers that approximates the properties of random numbers. The sequence is not truly random in that it is completely determined by a relatively small set of initial values, called the *PRNG*'s state, which includes a truly random seed. Although sequences that are closer to truly random can be generated using hardware random number generators, pseudorandom numbers are important in practice for their speed in number generation and their reproducibility.

Algorithm 4: PSEUDO RANDOM

G = (i, j) = mod [p(i, j) + e(i, j) , 256]

N1=row,N2=column
e.g.:100x200
N1=100,
N2=200

STEP 3: ENCRYPTION

Encryption is a common technique to uphold image security.An image can be grasped and data can be retrieved if it is in original form. Hence Block Based transformation algorithm is used to encrypt confidentially.

Algorithm 5: BLOCK BASED TRANSFORMATION

```
While I < NoBlocks
R = RandomNum between (zero and NoBlocks -1)
If R is not selected Then
Assign location R to the block I
I +=1
Else
If SEEDALTERNATE = 1 Then
seed = seed + (HashValue1 Mod I) +1
SEEDALTERNATE = 2
Else
seed = seed + (HashValue2 Mod I) + 1
SEEDALTERNATE = 1
Randomize (seed)
End If
Else
Number-of-seed-changes += 1
If Number-of-seed-changes > 500,000 then
For K = 0 to NoBlocks -1
If K not selected then
Assign location K to Block I
I=I+1
End if
Next K
End if
End if
```

STEP 4: TRANSMISSION

The encrypted image is transmitted to the receiver. The keys that are responsible for the retrieval of image is to be sent to the receiver. The image and the hidden data could be retrieved only if all the four keys were properly matched.

STEP 5: RETRIEVING ORIGINAL DATA

A content owner encrypts the original uncompressed image using an encryption key. Then by using least significant bits of the encrypted image is compressed. A data-hiding key is used to create a space to save some confidential information. If a receiver has the data-hiding key, then the image content can be retrieved. With the encryption key one can retrieve the image and not the confidential information. Data hiding key and encryption allowsa a user to retrieve both the original image and the confidential information. The data is not get lost by the authorized user by RDH at the corners.

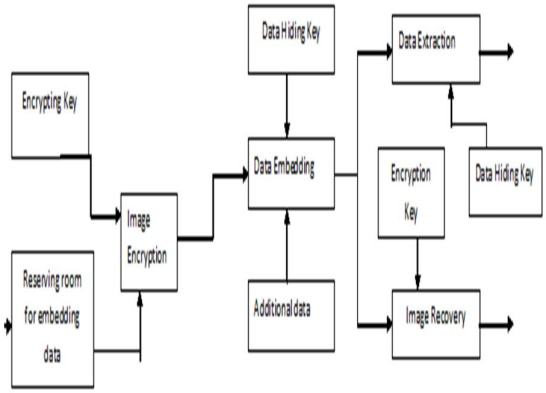

Figure 4: Proposed Architecture

Algorithm 6: RDH

```
pictureBox1.Image=Image.FromFile(EnImage_tbx.Text) ;
 if (saveFileDialog1.ShowDialog() == DialogResult.OK)
{
 saveToImage = saveFileDialog1.FileName;
 }
 else
  return;
  if (EnImage_tbx.Text == String.Empty || EnFile_tbx.Text
== String.Empty)
  {
   MessageBox.Show("Encryption information is
```

5. CONCLUSION

This project has proposed a novel scheme of scalable coding for stegno images. The data that get hidden in the image can be extracted by the intruders by using various techniques. This project used the various techniques like RDH,IWT,DCT,HBC to secure the data from the intruders. The Stegnanalysis methods can be used to retrieve the original data from the sender and the user can view the same quality of the stegno image as the original imgae has. The quality and the size is get maintained in this project. The HTML embedding can be used in the further future enhancement.

6. REFERENCES

[1] Katzenbeisser, S. and Petitcolas, F.A.P., (2000) Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, Inc., Boston, London

[2] M. F. Tolba, M. A. Ghonemy, I. A. Taha, A. S. Khalifa, (2004) "Using Integer Wavelet Transforms in Colored Image-Stegnography", International Journal on Intelligent Cooperative Information Systems, Volume 4, pp. 75-.

[3] Guorong Xuan et. al, (2002) “Distortionless Data Hiding Based on Integer Wavelet Transform”, Electronics Lelters, Vol. 38, No. 25, pp. 1646-1648.

[4] Shejul, A. A., Kulkarni, U.L., (2011) “A Secure Skin Tone based Steganography (SSTS) using Wavelet Transform”, International Journal of Computer Theory and Engineering, Vol.3

[5] Masud, Karim S.M., Rahman, M.S., Hossain, M.I., (2011) “A New Approach for LSB Based Image Steganography using Secret Key.”, Proceedings of 14th International Conference on Computer and Information Technology, IEEE Conference Publications, pp 286 – 291

[6]] Xie, Qing., Xie, Jianquan., Xiao, Yunhua., (2010) “A High Capacity Information Hiding Algorithm in Color Image.”, Proceedings of 2nd International Conference on E-Business and Information System Security, IEEE Conference Publications, pp 1-4.

[7] Sachdeva, S and Kumar, A., (2012) “Colour Image Steganography Based on Modified Quantization Table.”, Proceedings of Second International Conference on Advanced Computing & Communication Technologies , IEEE Conference Publications, pp 309 – 313.

[8] Chen, R. J., Peng, Y. C., Lin, J. J., Lai, J. L., Horng, S. J. Novel Multi-bit Bitwise Adaptive Embedding Algorithms with Minimum Error for Data Hiding. In Proceedings of 2010 Fourth International Conference on Network and System Security (NSS 2010), (Melbourne, Australia, 1-3 September 2010), IEEE Conference Publications, 306 – 311.

[9] Roy, S., Parekh, R., (2011) “A Secure Keyless Image Steganography Approach for Lossless RGB Images.”, Proceedings of International Conference on Communication, Computing & Security, ACM Publications, 573-576.

[10]] Mandal, J.K., Sengupta, M., (2011) “Steganographic Technique Based on Minimum Deviation of Fidelity (STMDF).”, Proceedings of Second International Conference on Emerging Applications of Information Technology, IEEE Conference Publications, pp 298 – 301

International Journal of Computer Applications Technology and Research
Volume 3– Issue 3, 140 - 145, 2014, ISSN: 2319–8656

An Interactive visual Textual Data Analysis by Event Detection and Extraction

Danu.R
Sri Manakula Vinayagar
Engineering College
Pudhucherry, India

Pradheep Narendran. P
Sri Manakula Vinayagar
Engineering College
Pudhucherry, India

Ranjith Kumar. C
Sri Manakula Vinayagar
Engineering College
Pudhucherry, India

Bharath. B
Sri Manakula Vinayagar
Engineering College
Pudhucherry, India

Abstract: Now a days, searching for the text data in a large ocean like location is quite challenging and more inaccurate task. Data that holds with the relation to its event can be evolved with certain changes with some intervals of time. Already existing techniques provides a trendy manner in order to extract a textual data with the visual analysis based on the event.But few data may have topic meaning that representing the kind of data to be extracted. In this paper, we propose a analytic system as an interactive manner called LeadLine, to recognize a data automatically by some semantic events in news blog as well as social media and deploys expansion or retrieval of the events. To organize such an events, LeadLine combines topic modeling, event detection, and named object or an entity recognition techniques to retrieve information automatically based on who, what, when, and where for each event. In order to make text data to be an effective one, LeadLine enables users to analyze interactively valid events by using 4 Ws to build an reviewing of mainly how, when and why. Bulky text data can be present normally as also the outdated one. These data can be concise with the help of LeadLine. LeadLine also provides the most simple process just by the exploration of events. To prove the effectiveness of LeadLine, These were implemented in the news blogs and social media data.

Keywords: Event detection, Topic modeling, LeadLine, Entity recognition.

1. INTRODUCTION

News blogs and online news like various text data present as a real-time dependent that is purely periodical based were located as worldwide. In the news blog, it has certain events that follows chain manner and in social media, the data can be simply like a user comments about something in the social aspect. Matching of certain patterns in terms of comprehensive can be either constant set of feed ot a changeable set of data. Some data in both the social media as well as news blog can be hidden in some case because of their privilege. So,a process to filter data that are in the form of text can be chosen based on their topics, and the relevant set of information can be triggered in order to get the assembling of complete appropriate information as a result. While examining the text data among the numerous amount of data, there will be more problem that can be faced from an event perspective .

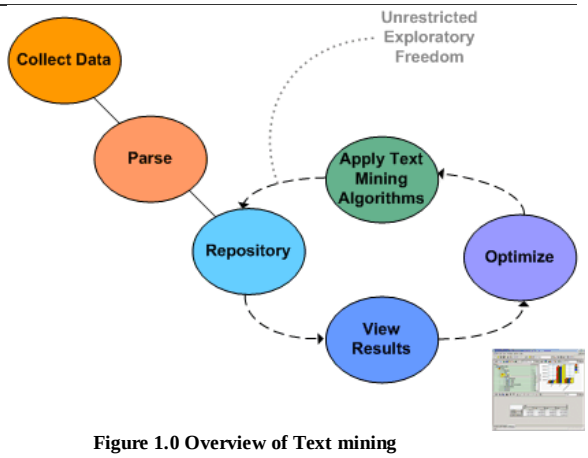

Figure 1.0 Overview of Text mining

There are many communities visualizing the working of the topic modeling with the time perspective. But here, we focusing on the topical trend based on the time, that doesn't meant for the complete event based technique but the major change in the temporal trends of the particular events.

1.1 LeadLine - An Introduction

A congested model that allows us to deploy computational methods to perform auto-extraction process for the events from text data. To explore such an events, we retrieve information based on what, who, where and when by simply integrating topic modeling, event detection, and entity recognition techniques. Initially, text data can be extracted in the social media and news blog sites on the conceptual themes using Latent Dirichlet Allocation (LDA) to provide topic in the formulation on events. To recognize the trending scale for each events, we have implemented an Early event detection algorithm to control the persistence of the events. This step of execution provides an attribute for representing the starting of any event that may also further expanded or depends upon other events as a ending event. To extract information about any people or location, related to the event, named entity recognition for the set of corpus of text and associate them with the events. With the above four processes are modeled in a system as an explicit one, our approach reinforces identification and extraction of events by topical, entity level and in trendy manner. To correlate and combine the events results as an effectively, we built a visual interface that suggests some related results for the event. Such an interface enables users to interactively traverse events and mainly to adjust or modify the event detection process based on the level of detailed set of data. Shaping the text data based on the event has additionally provides a base line for building such ideas as a creative. We have extended LeadLine that has a capacity to validate data, which allows its user to access and revisit the extracted findings easily. Especially, our approach provides three different benefits:

- Provides creative examining interface that makes users to get back their findings.
- A common process that integrates topic modeling, entity recognition, and already existing event detection mechanisms to identify semantic events from text data.
- An interactive visual system for analyzing user searchable textual data in the forms of 4W's set of questions.

1.2 Formulating Events

There are several questions that leads to critics to identify an event from the collection of text. How such meaningful events are carried out and extracted from the bulky collection of text ? Several properties that describes the characteristics of a specific event ? How to explore an event that in turn automatically discovers an appropriate event from the text corpora ? To reply these questions, we first make sure on what made up of an event:

Merrian - Webster defines a general definition that an event is a thing that happens or takes place, especially one of importance or any activity. In Topic Detection and Tracking (TDT) community and event detection [7,11], an event is defined based on its property as " a notation of something that represents the certain thing with corresponding time, topic and location on where it is associated ". Similarly the story telling concepts by McKee defines that an event refers to "creates semantic change in the temporal situation of a particular character" [13].

By integrating all these definitions about an event, is an " Occurrence reflecting any change in the larger amount of text data that utilizes the related topics at a specific time. This is defined in terms of topic and time, and related with the entities like an individual/ group of person and location ". We refers events with a four attributes like < Topic, Time, People, Location >. These refers to the 4W's questions : what, who, where and when.

2. RELATED WORKS

We mainly concentrate on the three areas such as named entity recognition, event detection, topic detection and analysis, and also text visualization techniques for a text with the background work of LeadLine.

2.1 Event Structure in Perception & Process

A different piece or the segment of time that denotes any person or location with the starting and ending stage is called as event. People can easily get them through the event just because of dividing and identify them with the different part of time continuously. People may use such an obeserved segments into an events or physical activities at mutiple set of timescales. Since, the same concept can be applied for even the abstract continuous streams, like topical streams, from the text corpora. Though, an event is treated the unit of making use ofactivities that serves more natural representation of any activities.

2.2 Event Detection

Over-the counter (OTC) medication sales, a type of as a source for detecting events indicating disease outbreaks describes a mutually growing system built for time detection of anthrax, a widespread occurrence of an infectious disease in a community at a particular time. This method comes into the category of common variation methods which concentrates on detecting events from time set of series [1]. As a more general approach, Guralnik et al. [2] presented steps to determine the change points in timely data dynamically without previous knowledge of the trending distributions. Other surveillance systems for a disease taken into consideration for both temporal and space related information. In addition, Neil at al. further developed a “multivariant Bayesian scanning statistic” (MBSS) [8] technique for fast and more relevant event detection. The already proposed event detection mechanisms are more efficient, but they lack the ability to handle text corpora,that may contain rich set of information that results with the symptoms and how they can be evlolved in the over time. In this paper, our approach allows to convert textual data into multiple semantic time information so that we can apply different ideas from the Biosurveillance community for early event detection on text data. The locality-sensitive hashing, enables first story detection on streaming data is chosen as a proposed system. However, the importance of the extracing events is not covered in the proposed technique.

2.3 Data Acquisition and Preparation

To explain the common techniques of our approach and their deployable domains, we have applied in two types of text data: CNN news and microblogs from Twitter. While both kind of origin that contain some collection of rich set of information resulting in a major real-world events, the main reason for choosing the two set of data sources is just because of its flexible editable module of style and the delay for responding to a particular event. In some specific, content from news media like CNN and others are customized by some journalists by specifying the topics with some set of background works. Not every but some posts in the social networks contains several information which are fixed range of commendatories [10]. These different type of text

information provide various levels of benchmarks that enables to validate the logical architecture.

3. SYSTEM ARCHITECTURE

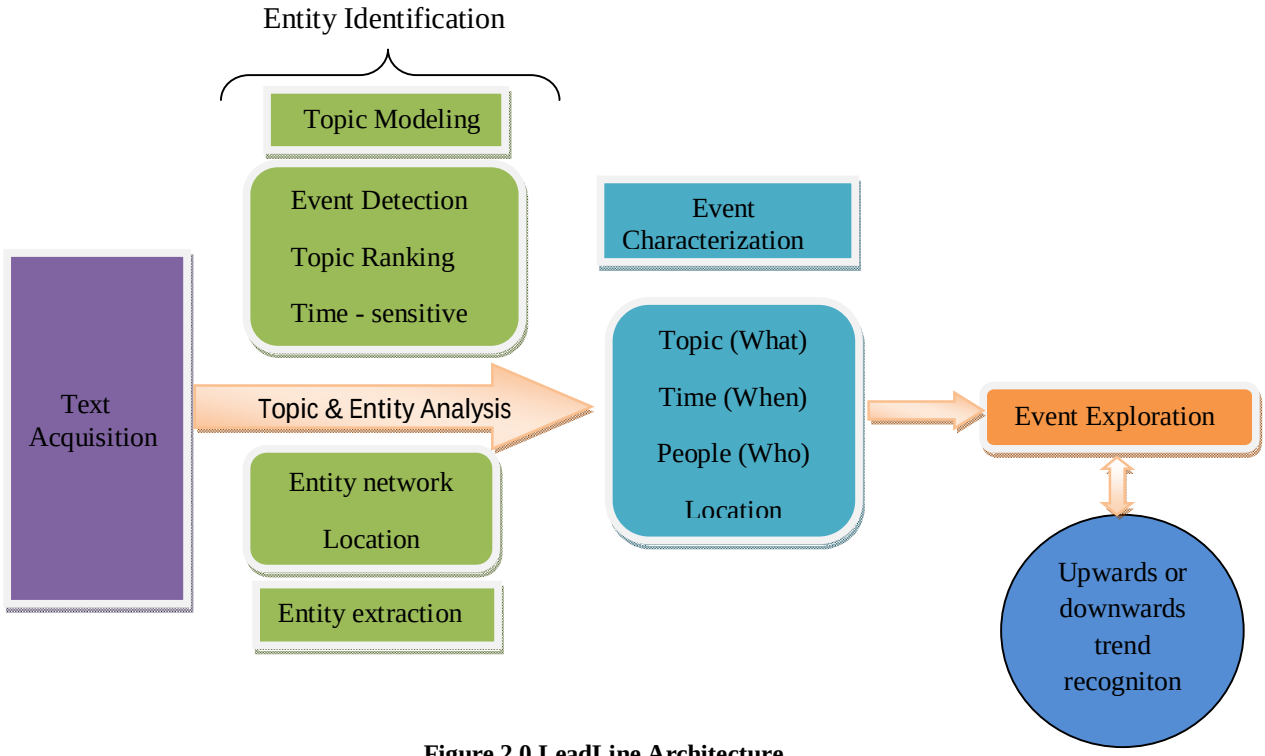

Figure 2.0 LeadLine Architecture

3.1 Data Acquisition and Preparation

To explain the common techniques of our approach and their deployable domains, we have applied in two types of text data: CNN news and microblogs from Twitter. While both kind of origin that contain some collection of rich set of information resulting in a major real-world events, the main reason for choosing the two set of data sources is just because of its flexible editable module of style and the delay for responding to a particular event. In some specific, content from news media like CNN and others are customized by some journalists by specifying the topics with some set of background works. Not every but some posts in the social networks contains several information which are fixed range of
commendatories [10]. These different type of text information provide various levels of benchmarks that enables to validate the logical architecture. These two sources belongs to public domain, there are no certain set of data that are not supposed available or visible as they are protected or under privacy enabled. Hence, we have extended our existing architecture in the news blogs and the social network with data crawling mechanism. The current approach extends the existing one is just by adding the news article crawling techniques. Both these news blog and the article data needs to be monitored and the extractor mechanism watches up to hour based set of tasks.

News Blog Data Acquisition: It needs to be customized with the page crawling and the RSS daemons, obviously. These methods generally implemented with universal techniques that tries to crawl the complete web domain information,copy all webpages, extract all the relevant textual articles, parse article time data. The data is stored into the HBase data structure that results in the faster access and MapReduce [9] based technique for the data cleaning and processing. Using these crawling techniques,data can be retrieved and filtered with the news articlesas the bottom up process.

Twitter Data Crawling: Some microblogs from Twitter, Facebook are also gathered in the form of dual crawling techniques. The primitive process that uses our MapReduce concurrently or a parallelized data crawler, which acts as between with the Internet through multiple

$$\sum_{v=1}^{N} (\sqrt{\beta_{i,v}} - \sqrt{\beta_{j,v}})^2 \qquad (1)$$

independent crawling techniques. Each crawlers may constantly gathers data from the social media by various public fields and moves it into HBase. As a result, we can able to collect over 5 billion posts or user tweets by providing a

reliable database from all languages over the course of 3 months for evaluation purposes. We implements a search technique called breadth-first search (BFS) using Nutch to obtain Twitter public user-graphs and capture them through their web portal for wider streams additionally.

3.2 Analytics architecture for events detection and characterization

To retrieve or extract an information from the text corpora, we can simply integrate the several kind of techniques to recongize <Time, People, Location, Topic>. To extract semantic topics with their timespan for any particular events, we holded topic models based on their themes and an Earlier Event Detection technique to identify a start and an end for each and every sort of event. To explore information about whom (individula person or a group) and where (associated location), we were performing named entity recognition and also analyzing relation between extracted data in the form of entities. We dividing the identification of topic themes and its span of time cycle as a topic-based way of analytics, in which we initially get through the topics from the input text corpora using Latent Dirichlet Allocation (LDA) as shown in the Fig 2.0. Then we applies, 1) topic - level event detection technique to automatically explore "events" as a triggers that are named by the timespan; 2) Time-tactful text or a phrase extraction that provides text information regarding an event with a set of brief keywords; 3) Topic ranking process to make easier of the discovery of event relation just by placing chunks of texts with similar topics nearby in a separate corpora; and finally Completing the topic-based analytics, our approach also focuses on named entity-based logic to identify people or/and a location relevant with each event. Especially, this process is interfaced as for extracting main/key entities from a textual data regarding whom and where. The visual interactive interface acts as a combining part of both logic processing to connect through the users and its complex analytic results. With this visual interactive interface, LeadLine mechanism supports interactive exploration of any events from various categories like whom, which, when, where as well as makes users to interact with the ongoing logical algorithms to partially makes adjusting the process of detecting and characterizing events from text corpora.

3.3 Topic-Based Event analysis and visual perception

Topic-based logic is a most crucial task in the event characterization in terms of exploring the topical theme and its time. Here, we just introduce an algorithm to extract topical and trendic information with based on an event, and some visual way of representations that can communicate with the topical as well as temporal ways.

3.4 Extracting Topics from Text Data

We begins by managing textual data streams depending on their topics. User simply gives a text as a word or a phrase, and there are different aspects to retrieve semantic topical themes. Among those aspects, Probabilistic topic models [12] are treated to be beneficious when comparing to traditional vector-based text analyzing techniques. In LeadLine, we first works with the most commonly used topic model called, LDA [14], to explore meaningful topics from text corpora.

5.1.1 Visualization of Topic Streams

To represent a data with the specialization of how it visually has to be presented, it merely concentrates on how the height of the topical themes that are changed in a searching domain. Each topic contains some relevant data information that can be carried out with the sort of holding some topical information about the searching data. Still, more effective algorithms are used, it wont results an exact crispy topical contents are retrieved in a system. In order to serve the complete context, a ThemeRiver representation is used in the backdrop of the visualization process. Thus redundant text patterns revealed by a text stream as a row (like weekly data pattern in the news stories) are still depicted.

5.1.2 Topic Streams

Time is more important attribute of an event than the topic. For making enabled of the clear process about the temporal change observing and exploring, we manage those topics along with the temporal central line. By considering each topic as a data information that exceeds over the time, the calculation of each topic information is done by processing a container based on the amount of text information related with the theme of the topic in each timescale. It is a unit that in which texts are integrated based on the temporal behavior. The time frame unit can be simply differs by collection of data and its tasks, that can be ranging also for minutes frequently in the social media data into days for newer stories.

3.5 Topic Ranking

During the exploration of any event, the results retrieved to be visually kept placed onto the similar set of visualization can be holded in a contiguous manner and also the events recently derived are topic-based ranked. LDA approach does not explicitly make the relationship between their corresponding topics, so that we need to rank the topics which is identical to be founded by Hellinger distance.

4. RESEARCH PROPOSAL

STEP 1: Automatically Detecting Events in Topical Streams

A major task of this approach is to detect the temporal changes that are happening to the event. To detect such events, based on the topical theme, we need to consider it with the help of time series. Each and every time series is computed by relating or aggregating each topic with its assigned time scale. Most probably, we use the cumulative sum control chart (CUSUM) for the purpose of change detection [15]. It is effective for recognizing variations in the mean in a time series by maintaining a running sum of "revelation". We adopted CUSUM maily for detecting changes in topical data theme. For every topic theme, the program keeps itself a mutual integration maintains the topical theme and each stream has its own time span that are high when comparing to mean topic. To automatically retrieve data information in the topic streams, the mechanism called Earlier Event Detection (EED) can be used to identify bursts to a particular event. If the mutually integrated sum is more than a threshold, the event can be triggered out. The result is a set of automatically detected events within all topic streams, with each event labeled by a start and an end along the time dimension. If the data can be expected for the future events are to be represented, then a file that contains relevant details between two dot operators are pulled out. Such a process of detecting timely events are more reliable task.

4.1 Visualizing Detected Events

To present any topical streams with the information as a visually interactive, we have an outline of its representation as well as highlighting the events of those particular topical stream which would have been chosen. The wider information of a time of the outlined data is choosen by the event detection resultant data. In addition to it, LeadLine mechanism supports starring of an events as a suggested or an interest via its user interaction process. To provide information in a crispy manner, LeadLine enables its users to access a documents like news feeds or microblog data can be defined as just by clicking on the event.

ALGORITHM 1 : Cumulative Check sum

Algorithm : CUSUM
Input:Various topical time series X collected $i = 1,...k$
Steps:
1. Calculate the mean μ and standard deviation σ of the particular time series;
2. Calculate presently running sum S from the starting time scale
S1 = max[0, x_1 - μ]
Si = max[0, S_{i-1} + x_i - μ].
3. When S_k exceeds a value exists in H (in units of σ), event triggers. The start and end of the event are determined by the closest positive Si to its triggering point.
4. If time is not mentioned, or any keywords like 'Upcoming', 'Future', 'List of any events' then
Date = Get (Today's date)
Explore all the topical data that consists of information within two dot operators that exceeds the Date.

4.2 Detection of an Interactive Event

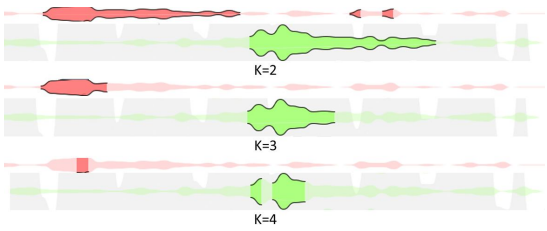

Figure 3.0 Comparison of different variations of the events.

One of the most striking advantage of this approach is just for providing automatically for an event detection that mainly triggers for the topical streams that are treated as a bursts which guarantees for the particular event. By simply clicking the button called as "Tune", the user can able to adjust the fine or coarse of the discovery . K refers to the standard deviation which are usually said to be a fixed mesasure of the threshold [16]. Users are allowed to adjust those K values . If the value of K is minimum, then there will be a situation of making sure that there are lesser number of mean of the variation on the particular event. If the value of K is greater as found, then there will a result of bigger range of shifts. If the user adjusts the tune level, then the LeadLine mechanism has to re-execute with the present values in the system.

STEP 2: Time-sensitive Keyword extraction

To make an approach an efficient one, we need to refine the search and more recent information has to be provided to the user. In order to perform such an operation, we need to provide each event with its own time scale based retrieval process. The input for this algorithm is a text data that can be divided into sub-collections using their time scale and also in topics. Each sub - collection of data may have its own timespan and the topic recognized entity. The algorithm follows a TF-IDF (**Term Frequency**–Inverse Document Frequency) heuristic to choose time-sensitive terms: (a) if a term occurs many times in the sub-collection, it is marked; (b) if the term also occurs in many of other set of sub-collections, the importance is not marked.

ALGORITHM 2: Extract time-sensitive terms

Input: Topic-term distribution matrix ϕ; desired number of keywords per time frame N
Steps:
1.for each topic i do
for each time frame t do
Identify a collection of documents $D_{i,t}$ focusing on topic i from entire text stream;
end for
end for
2. for each term W in topic i from $D_{i,t}$ do
calculate term frequencies Time Frequency
end for
3. Re-rank the Time Frequency scores with topic-term probabilities[17]
4. Within each topic and time frame, select the top N terms astime-sensitive terms.

5. CONCLUSION

In this paper, we were enhancing an visual interactive analytics system called as LeadLine, that identifies semantic events and enables users to validate the changes in the social media as well as news feeds topical streams from the triggering of events. To explore such an events, LeadLine mechanism uses who, what, when and where conditions to retrieve information based on the categories. It also provides a visually interactive process in a system. Finally, the results obtained by LeadLine doesn't only have semantic information, but also provides its user a complete information about his data.

6. REFERENCES

[1] A. Goldenberg, G. Shmueli, R. A. Caruana, and S. E. Fienberg. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proceedings of the National Academy of Sciences of the United States of America, 99(8):pp. 5237–5240, 2002.

[2] V. Guralnik and J. Srivastava. Event detection from time series data. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '99, pages 33–42, New York, NY, USA, 1999. ACM.

[3] J. Allan, editor. Topic detection and tracking: event-based information organization. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[4] D. Neill and G. Cooper. A multivariate bayesian scan statistic for early event detection and characterization. Machine Learning.

[5] Apache hadoop. http://hadoop.apache.org, 2012.

[6] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. Eddi: interactive topic-based browsing of social status streams. In Proceedings of the 23nd annual ACM symposium on User interface software and technology, UIST '10, pages 303–312, New York, NY, USA, 2010. ACM.

[7] H. Mannila, H. Toivonen, and A. Inkeri Verkamo. Discovery of frequent episodes in event sequences. Data Min. Knowl. Discov., 1(3):259–289, Jan. 1997.

[8] D. Blei and J. Lafferty. Text Mining: Theory and Applications, chapter Topic Models. Taylor and Francis, 2009.

[9] R. Mckee. Story - Substance, Structure, Style, and the Principles of Screenwriting. Methuen, 1999.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, March 2003.

[11] D. C. Montgomery. Statistical quality control. Wiley Hoboken, N.J., 2009.

[12] D. B. Neill and W.-K. Wong. Tutorial on event detection tutorial. http://www.cs.cmu.edu/ neill/papers/eventdetection.pdf, 2009.

[13] LeadLine: Interactive Visual Analysis of Text Data through Event Identification and Exploration, IEEE Conference on Visual Analytics Science and Technology 2012.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 3, 146 - 149, 2014, ISSN: 2319–8656

Mobile Device Protection Using Sensors

Anna Rose Vembil
Department of Computer Science and Engineering
Jyothi Engineering College, Cheruthuruthy, Thrissur, India.

Shilna Latheef
Department of Computer Science and Engineering
Jyothi Engineering College, Cheruthuruthy, Thrissur, India

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

Anil Antony
Department of Computer Science and Engineering
Jyothi Engineering College, Cheruthuruthy, Thrissur, India

Abstract: Mobile devices like laptops, iPhones and PDAs are highly susceptible to theft in public places like airport terminal, library and cafe. Moreover, the exposure of sensitive data stored in the mobile device could be more damaging than the loss of device itself. In this work, we propose and implement a mobile device protection system using sensors, based on sensing and wireless networking technologies. Comparing with existing solutions, it is unique in providing an integrated protection to both device and data. It is a context-aware system which adjusts the protection level to the mobile device dynamically according to the context information such as the user proximity to the mobile device, which is collected via the interactions between the sensors carried by the user, embedded with the mobile device and deployed in the surrounding environment.
Keywords: User Sensor (US), Mobile Device Sensor (MDS), Advanced Encryption Standard (AES), Central Server (CS)

1. INTRODUCTION

Mobile devices, such as laptops, smart phones and PDAs, have become an essential part of our daily life. They are small and easy to carry but also powerful in computational and storage capabilities. Unfortunately, these merits also put them at risk. For example, because mobile devices are small, they usually are highly susceptible to theft, especially at public places like airport terminal, library and cafe. As mobile devices get slimmer and more powerful, the number of mobile device thefts surges.

On the other hand, keeping data secure in a mobile device is a critical requirement. Unfortunately, a majority of the mobile device users do not take necessary actions to protect the data stored in their mobile devices. Therefore, the loss of a mobile device could mean the loss and exposure of sensitive information stored in the lost device, which may be much more valuable than the device itself. In this paper, we propose a mobile device protection system for sensors, with the help from sensing and wireless networking technologies. We deploy low-cost wireless devices at public places of our interest. Users and mobile devices carry special-purpose wireless sensing devices which provide protection to the mobile device and the data stored in it.

Specifically, this paper has the following unique features:

• Context Awareness: Sensors carried by the user and the mobile device interact with each other to collect context information (e.g., proximity of the user to the mobile device) and then the system adapts its behavior properly and promptly to the context change.

– Anti-theft Protection for Mobile Device: When the user is away from the mobile device, system monitors the mobile device. When a potential theft is detected, system quickly alerts the user.

– Data Protection: System adapts the protection level for data stored in the mobile device and incorporates a carefully-designed authentication mechanism to eliminate possible security attacks.

• Low-cost and Light-weight: System utilizes low-cost sensors and networking devices. The software implementation is light-weight and may be adapted for mobile devices of various kinds.

2. RELATED WORKS

a. Mobile Device Protection

Different models exist for the protection of the mobile device against theft In general, they can be classified into the following two categories: *recovery/tracking-oriented systems* and *prevention-oriented systems* .In *recovery/tracking –oriented* systems, a back-end software process runs on the device, which can send "help" messages across the Internet to the tracking service provider in case the device is lost or stolen. The service provider can pinpoint the location of the lost device based on the "help" messages. These systems are ineffective in preventing mobile device thefts since they aim at recovering the devices at theft.

In comparison, prevention-oriented systems aim at deterring the adversary from compromising the mobile device. When a potential theft is detected, the system raises an audible alarm to deter the adversary from completing the theft. Ka Yang, Nalin Subramanian, Daji Qiao, and Wensheng Zhang proposed a context-aware system which adjusts the protection level to the mobile device dynamically according to the context information such as the user proximity to the mobile device, which is collected via the interactions between the sensors carried by the user, embedded with the mobile device and deployed in the surrounding environment. When a potential theft is detected, an audible alarm will be triggered to deter the adversary from completing the theft. At the same time, alert messages will also be sent to the user. The MDS initiates the alert messages and sends them either directly to the user if the user is within direct communication range to the mobile device, or via the wireless network infrastructure. [1]

b. Data Protection

There are systems which give importance to the protection of the data. Mark D Corner and Brian D. Noble, proposed a system, where the user wears a small authentication token that communicates with a laptop over a short-range, wireless link.Whenever the laptop needs decryption authority, it acquires it from the token. The token will continuously authenticate to the laptop by means of a short-range, wireless link. Each on-disk object is encrypted by some symmetric key, Ke. File decryption takes place on the laptop, not the token. The file system stores each Ke, encrypted by some key-encrypting key, Kk. Only tokens know key-encrypting keys. A token with the appropriate Kk can decrypt Ke, and hence able to read files encrypted by Ke[2]. Carl E. Landwehr proposed a system where the user is given a token called wireless identification agent. It consists of a key unique to that WIA to each user. Once per *Tre-ident*or when prompted by the WIA, the user enters the PIN, becoming an

identified user. The detector attached to the workstation verifies the user. If the Detector fails to get a valid response from the current user's WIA within specified period *Td*, the Detector blanks the screen and disables the keyboard. Thus prevents the thief from accessing the data[4].Eagle vision protects the data the file system on the mobile device by encryption using a symmetric key *K*enc , which allows lower encryption and decryption latency. *K*enc is protected with a PKI public key *K*pub and the encrypted symmetric key *{K*enc*}K*pub is stored on the mobile device [1].

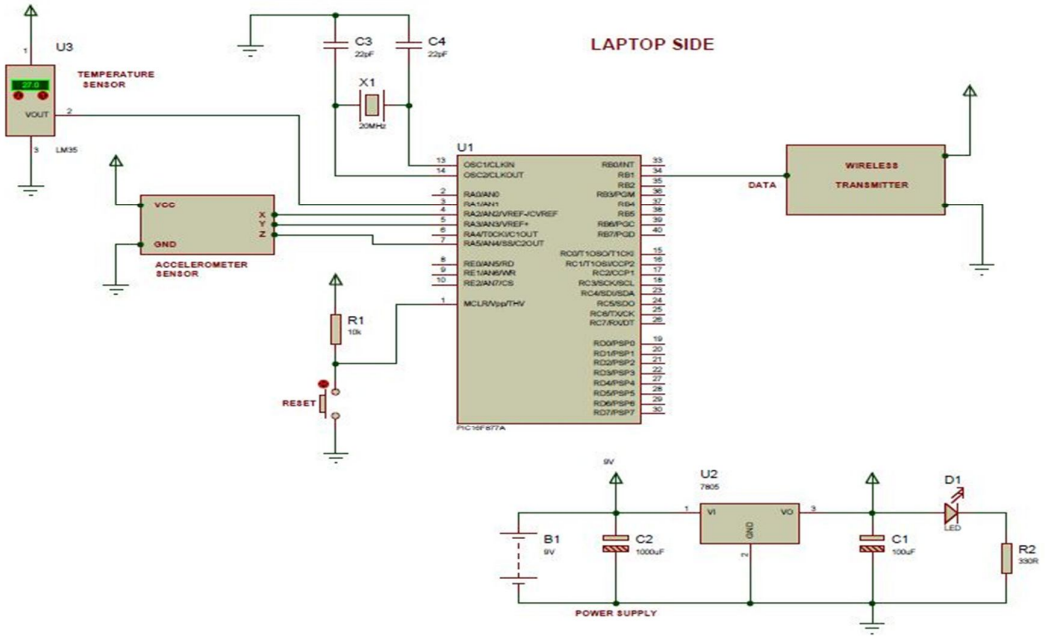

Figure 1 Mobile Sensor Circuit Diagram

3. PROPOSED SYSTEM

In this section we discuss about our proposed system. The proposed scheme enhances the security level in our mobile device by providing various features.

A. *Temperature Detection*

When the temperature of the surrounding environment increases it can cause damage to the mobile devices. Here we set a certain threshold value of temperature for the mobile device. When the value of the temperature exceeds this threshold value, an alert is send to the user by setting an alarm and the important data that is selected by the user is stored as a backup. The temperature sensor we used here is LM35.

B. *Low Cost*

In other papers a two way communication is maintained which involves the use of ZigBee which is very costly. But in this paper we implement a one way communication between the US and the MDS which does not involve the use of such expensive sensors and thus makes it a cost effective system.

C. *Encryption and File Transfer*

When a threat is detected the files in the mobile device are encrypted and sent to the central server. The files are selected according to their importance by the user. The files are transferred to the server through socket programming.

D. *Alert*

The mobile device sensor consists of an accelerometer having x, y and z coordinates. A certain threshold value for these coordinates is defined. When the value of these coordinates exceeds this threshold value it gives an alert to the user by setting an alarm.

4. IMPLEMENTATION

We demonstrate our mobile device protection system for the following features: anti-theft protection, privacy protection,

alerts dispatch and context awareness. In the following, we present a) details of hardware b) system model, c) trust and threat model and d) an example scenario.

a. Hardware Components

We implement the mobile device protection system using various components. We have a mobile device sensor as well as user sensor.. Mobile device sensor and user sensor communicate with each other using RF module. Mobile device sensor consists of RF transmitter and user sensor consists of RF receiver. As the laptop moves the accelerometer detects the motion. We use PIC16F877A here. The change in y, z coordinate and the temperature is noted. The temperature sensor used here is LM35. We use an encoder HT12E in mobile device sensor which encodes the value and passes it to the transmitter.

Transmitter sends this value to the receiver. We use a decoder HT12D at user sensor which decodes the value received from receiver. We use RS232 in mobile device sensor. RS232 is the traditional name for a series of standards for serial binary single ended data and control signals connecting between DTE (data terminal equipment) and DCE (data circuit-terminating equipment). It is commonly used in computer serial ports. We also use MAX232

that converts signals from an RS-232 serial port to signals suitable for use in TTL compatible digital logic circuits.

We also use 7805 regulator in mobile device sensor. 7805 is a voltage regulator integrated circuit. The voltage source in a circuit may have fluctuations and could not give the fixed voltage output. The voltage regulator maintains the output voltage at a constant value. Capacitors of suitable values can be connected at input and output pins depending upon the respective voltage levels. An accelerometer is a device that measures proper acceleration. Here we use MMA7260 sensor. 3-Axis accelerometer with selectable range and low-power sleep mode. The MMA7260Q from Free scale is a very nice sensor with easy analog interface. Runs at 3.3V with 3 analog output channels for the three axes. An accelerometer is a sensor that measures the physical acceleration experienced by an object due to inertial forces or due to mechanical excitation. Acceleration is defined as rate of change of velocity with respect to time. It is a measure of how fast speed changes. It is a vector quantity having both magnitude and direction. As a speedometer is a meter to measures speed, an accelerometer is a meter to measure acceleration. An ability of an accelerometer to sense acceleration can be put to use to measure a variety of things like tilt, vibration, rotation, collision, gravity, etc.

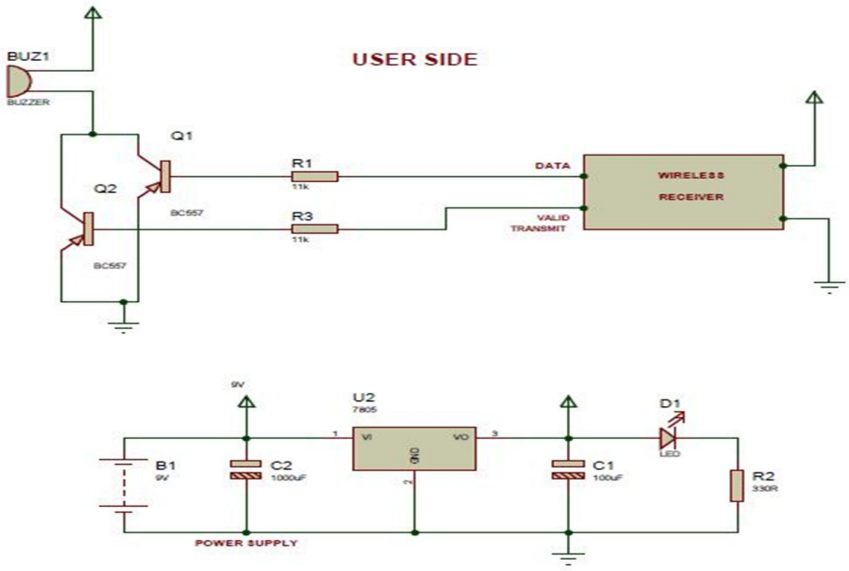

Figure 2 User Sensor Circuit Diagram

b. *System Model*

The Mobile Device protection System using Sensors consists of three components: *Mobile Device Sensor (MDS)*, *User Sensor (US)* and *Central Server (CS)*. Each mobile device carries an MDS which has several embedded sensors like an accelerometer and a temperature sensor. The MDS can communicate wirelessly with other system components. User of the mobile device carries a US, which interacts with other system components.

The accelerometer in the MDS detects the motion of the device and the temperature sensor present detects the temperature of the surrounding atmosphere. The MDS constantly interacts with the US using RF transmitter and receiver. This has a unique ID which helps in identifying the User Sensor the CS keeps the information about users and their mobile devices.

c. *Trust and Threat Model*

In our project the CS is considered to be trustable. A US is assumed to be secure as long as it is in the user's possession. An MDS is assumed to be secure when the user is nearby but may be tampered by the adversary if the user is away.

d. *An Example Scenario*

The following example scenario explains how our project works. Suppose Alice enters a library reading room with her laptop. Alice sets priority to certain files and Alice leaves the reading room to get some coffee from the café. Laptop's MDS starts to sample its accelerometer to detect any movement of the laptop and the temperature sensor checks the surrounding temperature for fluctuations. If a sudden movement is detected or the temperature rises, the laptop's MDS triggers an alarm in the US and the prioritized files are sent to the Central Server. Also the monitor locks itself. Alice can then decrypt files from the Central Server. The working is as given in figure 3.

5. CONCLUSION

In this paper, we propose a mobile device protection system. It is a context-aware system and protects the data stored in the mobile device in an adaptive manner .We implements this system using a mobile device, which consist of an accelerometer and a temperature sensor. It detects the motion as well as temperature. As motion is detected, the files are encrypted and transferred to the server. This system responds promptly to the context change and provides adequate protection to data, while not requiring explicit user intervention or causing extra distractions to the user. Future work includes further improvement of the system responsiveness.

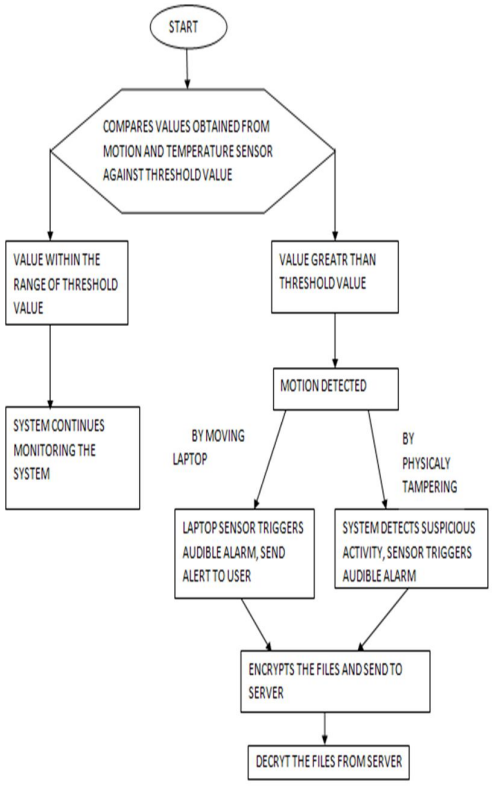

Figure 3 An Example Scenario

6. REFERENCES

[1] Eagle Vision: A Pervasive Mobile Device Protection System Ka Yang, Nalin Subramanian, DajiQiao, and Wensheng Zhang Iowa State University, Ames, Iowa – 50011

[2] M. D. Corner and B. D. Noble, "Zero-interaction authentication," in Proceedings of the 8th annual international conference on Mobile computing And networking, 2002.

[3] Mobile Device Security Using Transient Authentication, Anthony J. Nicholson, Mark D. Corner, and Brian D. Noble.

[4] Protecting Unattended Computers without Software, Carl E. Landwehr Naval Research Laboratory Code 5542Washington DC 20375-5337.

[5] Self Encryption Scheme for Data Protection in Mobile Devices, Yu Chen and Wei-Shinn Ku, Dept. of Electrical and Computer Engineering, SUNY Binghamton, Binghamton, NY 13902, Dept. of Computer Science and Software Engineering Aubum University, Aubum, Aubum AL 36849.

[6] A Hardware Implementation of Advanced Encryption Standard (AES) Algorithm using System Verilog. Bahram Hakhamaneshi, B. S. Islamic Azad University, Iran, 2004.

Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park

International Journal of Computer Applications Technology and Research
Volume 3– Issue 3, 150 - 154, 2014, ISSN: 2319–8656

Brain Tumor Detection Using Artificial Neural Network Fuzzy Inference System (ANFIS)

R. J.Deshmukh
Matoshri College of Engineering and Research Center
Nasik, India.

R.S Khule
Matoshri College of Engineering and Research Center
Nasik, India

Abstract: Manual classification of brain tumor is time devastating and bestows ambiguous results. Automatic image classification is emergent thriving research area in medical field. In the proposed methodology, features are extracted from raw images which are then fed to ANFIS (Artificial neural fuzzy inference system).ANFIS being neuro-fuzzy system harness power of both hence it proves to be a sophisticated framework for multiobject classification. A comprehensive feature set and fuzzy rules are selected to classify an abnormal image to the corresponding tumor type. This proposed technique is fast in execution, efficient in classification and easy in implementation.

Keywords: EEG**;** GLCM; ANFIS; FIS;BPN

1. INTRODUCTION

Manual brain tumor detection is time consuming and bestows ambiguous classification. Hence, there is need for automated classification of brain tumor. Normally, this turnover takes place in an orderly and controlled manner. The cells of tumor continue to separate, developing into a lump, which is called a tumor. brain tumor is divided in two types, primary and secondary brain tumor. The recognition of primary brain tumor is possible by observing the EEG (Electroencephalography) signals. EEG has been used to render a clearer overall view of the brain functioning at initial diagnosis stages. Being a non-invasive low cost procedure, the EEG is an attractive tumor diagnosis method on its own. It is a reliable tool for the glioma tumor series. The EEG in vascular lesions shows abnormality on first instance where as a CT scan shows abnormal on the third or fourth day .Medical Resonance images include a noise which is created due to operator's method of detection which can lead to serious inaccuracies in classification of brain tumor [1]. With increasing problems of brain, it is vital to develop a system with novel algorithms to detect brain tumor efficiently .The present method detects tumor area by darkening the tumor portion and enhances the image for detection of other brain diseases in human being. A comprehensive feature set and fuzzy rules are selected to classify an abnormal image to the corresponding tumor. Section I explores introduction of previous implemented techniques, Section II presents research work, Section III proposes the methodology used Section IV shows the simulation results, and Section V gives the conclusion.

The author in [2] employed the Hidden Markov Random Field (HMRF) for segmentation of Brain MRI by using Expectation-Maximization algorithm. The study shows that HMRF can be merged with other techniques with ease. The proposed technique acts as a general method that can be applied to a range of image segmentation problems with improved results.

Ahmed [3] explored an customized algorithm used for estimation of intensity of homogeneity using fuzzy logic that supports fuzzy segmentation of MRI data. The proposed algorithm is articulated by altering the objective function used in the standard FCM algorithm.

Habl, M. and Bauer, Ch. and Ziegaus, Ch., Lang, Elmar and Schulmeyer, F [4] presented a technique to detect and characterize brain tumors. They removed location arifactual signals, applied a flexible ICA algorithm which does not rely on a priori assumptions about unknown source distribution. Author have shown that tumor related EEG signals can be isolated into single independent ICA components. Such signals where not observed in EEG trace of normal patients.

2. PROPOSED METHODOLOGY

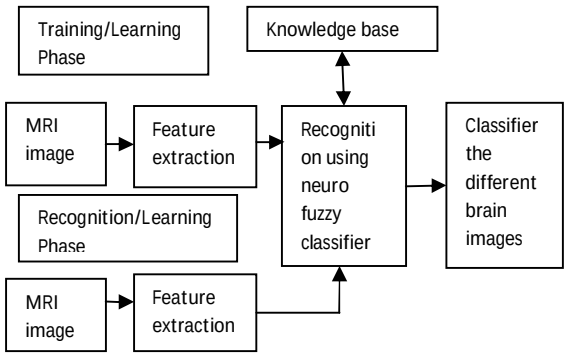

Figure.1 Block diagram of proposed system

Neuro-fuzzy systems use the combined power of two methods: fuzzy logic and artificial neural network (ANN). This type of hybrid system called as ANFIS ensures detection of the tumor in the input MRI image . The work carried out involves processing of MRI images of brain cancer affected patients for detection and Classification on different types of brain tumors. A suitable Nero Fuzzy classifier is developed to recognize the different types of brain tumors. Steps which are carried out for detection of tumor is enlisted below.

Step 1: Consider MRI scan image of brain of patients.
Step 3: Train the neural network with database images.
Step 2: Test MRI scan with the knowledge base.
Step 3: Two cases will come forward.
i. Tumor detected

ii. Tumor not detected.

A. Database Preparation:

The brain MRI images consisting of malignant and benign tumors were collected from open source database and some hospitals.

B. Image Segmentation:
The main objective of segmentation is to detach the tumor from its background.

C. Histogram Equalization:

The histogram of an image represents the relative frequency of occurrences of the various gray levels in the image.. Histogram equalization employs a monotonic, non-linear mapping which re-assign the intensity values of pixels in the input image such that the output image contains a uniform distribution of intensities.

D. Sharpening Filter

Sharpening filters work by increasing contrast at edges to highlight fine detail or enhance detail that has been blurred.

E. Feature Extraction*:*
The feature extraction extracts the features of importance for image recognition. The feature extracted gives the property of the text character, which can be used for training in the database. The obtained trained feature is compared with the test sample feature obtained and classified as one of the extracted character.

2.1 Feed Forward Neural Network

Figure 3 demonstrates the strategy of the Feed Forward for detecting the existence of the tumor in the input MRI. image, which is accomplished in the final categorization step. Here we use the Feed Forward neural network classifier to classify the image.

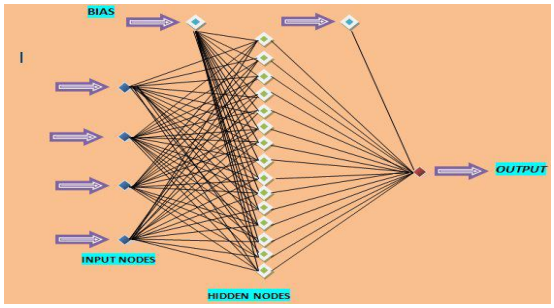

Figure. 2 Feed forward neural networks

Figure.3.Depicting back-propagation learning rule which can be used to adjust the weights and biases of networks to minimize the sum squared error of the network [8].

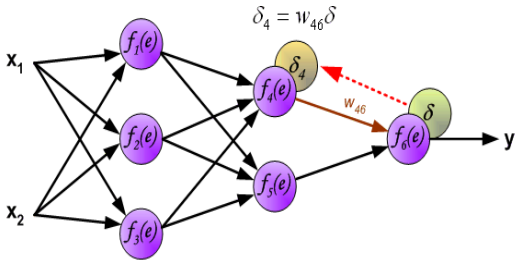

Figure. 3.Depicts the flow of information from output node back to hidden layer to reduce error.

The activation function considered for each node in the network is the binary sigmoid function defined (sgn = 1) as output = $1/(1+e^{-x})$, where x is the sum of the weighted inputs to that particular node. This is a common function used in many BPN. This function limits the output of all nodes in the network to be between 0 and 1. Neural networks are basically trained until the error for each training iteration stops decreasing. The features which are extracted from image are listed below.

Angular second moment:

$$f_1 = \sum_i \sum_j P(i,j)^2 \quad (1)$$

Contrast:

$$f_2 = \sum_{n=0}^{N_g-1} n^2 P_{x-y}(n) \quad (2)$$

Correlation:

$$f_3 = \frac{\sum_i \sum_j (ij) P(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (3)$$

Sum of Square: Variance:

$$f_4 = \sum_i \sum_j (i-\mu)^2 \quad (4)$$

Inverse difference moment:

$$f_5 = \sum_i \sum_j \frac{1}{1+(i-j)^2} P(i-j) \quad (5)$$

Sum Average:

$$f_6 = \sum_{i=2}^{2N_g} i P_{x+y}(i) \quad (6)$$

Sum Variance:

$$f_7 = \sum_{i=2}^{2N_g} (i-f_6)^2 P_{x+y}(i) \quad (7)$$

Sum entropy:

$$f_8 = -\sum_i^{2N_g} P_{x+y}(i) \log P_{x+y}(i) \quad (8)$$

Entropy:

$$f_9 = -\sum_i \sum_j P(i,j) \log P(i,j) \quad (9)$$

Difference variance:

$$f_{10} = -\sum_{i=0}^{N_g-1} (i - \mu_x - y)^2 P_{x-y}(i) \quad (10)$$

Difference entropy:

$$f_{11} = -\sum_{i=0}^{N_g-1} P_{x-y}(i) \log(P_{x-y}(i)) \quad (11)$$

Standard deviation

$$f_{12} = \frac{\sum_{i=1}^{n} (xi - \bar{x})^2}{(n-1)} \quad (12)$$

Where P(i,j) is $(i,j)^{th}$ entry in a normalized gray-tone spatial-dependence matrix.
$P_x(i)$ is i^{th} entry in the marginal-probability matrix obtained by summing the rows of P(i,j),= $\sum_j^N g, P(i,j)$.
N_g Number of distinct gray levels in the quantized image.
μ_x, μ_y, σ_x, and σ_y are the measured standard deviations of P_x, P_y.

3. GUI OF PROPOSED SYSTEM

Figure. 5 shows the GUI of proposed system for brain tumor detection. Figure. 6 showing the database of images containing tumor.Fig7 showing histogram equalization of input image in which intensity of image are equalized. Figure 8 showing segmentation of image in which tumor part is isolated from background. Figure 9 showing feature extraction of input image containing tumor.Figure.11showing detection of tumor

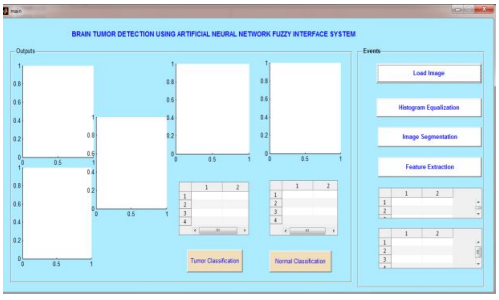

Figure. 5 Screenshot of GUI

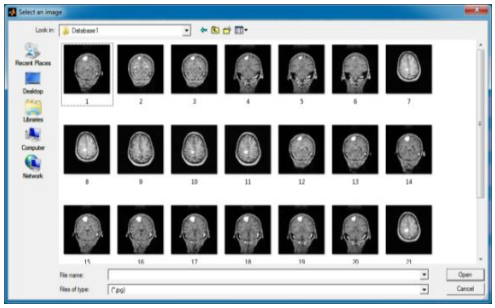

Figure.6 Screenshot images of brain tumor

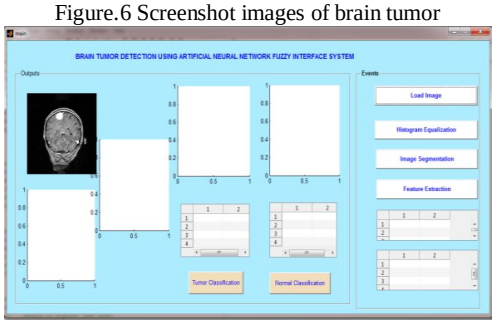

Figure.7 Screenshot showing loading of MRI image

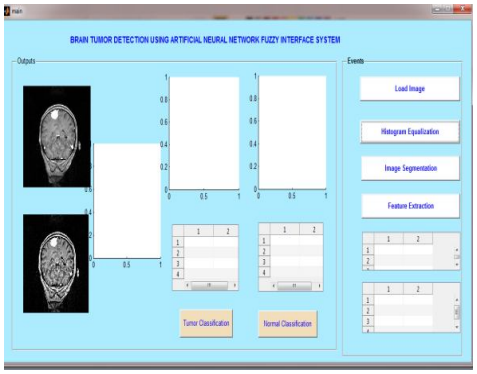

Figure.8 Screenshot showing histogram equalization

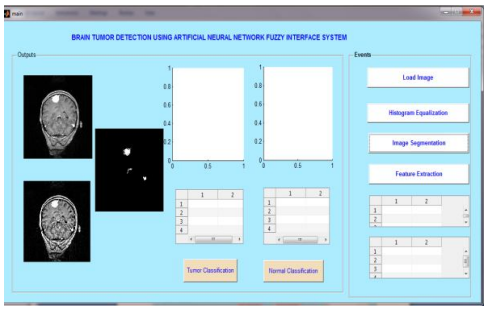

Figure.9 Screenshot showing image segmentation

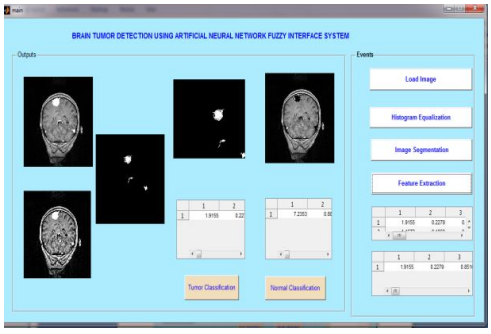

Figure.10 Screenshot showing feature extraction.

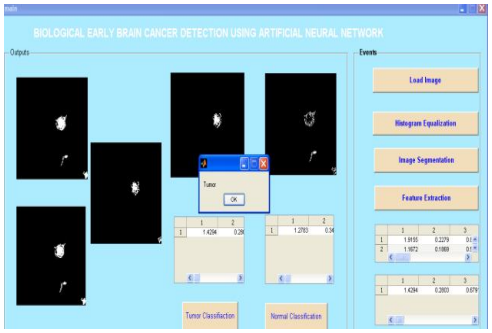

Figure. 11 Screenshot showing detection of tumor

4. NEURO-FUZZY CLASSIFIER

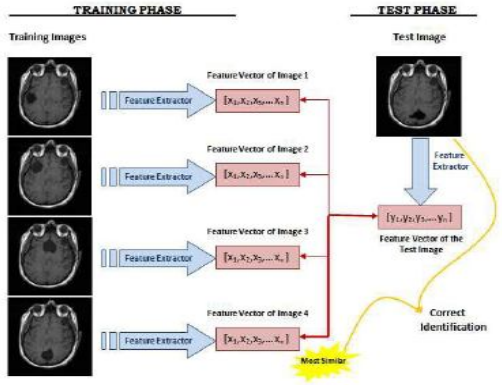

Figure.12 Testing and training phase of ANFIS

The features extracted from image are further given to Neuro-fuzzy classifier which is used to detect candidate circumscribed tumor. Generally, the input layer consists of seven neurons corresponding to the seven features. The output layer consists of one neuron indicating whether the MRI is a candidate circumscribed tumor or not, and the hidden layer changes according to the number of rules that give best recognition rate for each group of features.

5. SIMULATION RESULTS

Fig 11 shows the GUI neural network toolbox.
Fig 12 shows Performance Plot mean square error dynamics for all your datasets in logarithmic scale. Training MSE is always decreasing with increasing in number of epochs.

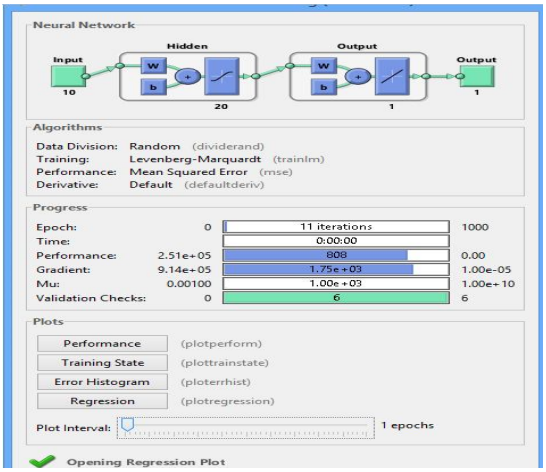

Figure. 13 Screenshot of GUI neural network training phase.

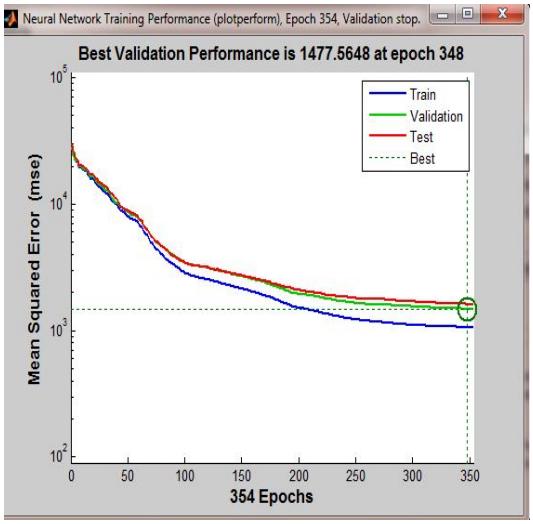

Figure.14 Screenshot of validation phase of neural network

6. CONCLUSION

This paper presents a automated recognition system for the MRI image using the neuro fuzzy logic. It is observed that the system result in better classification during the recognition process. The considerable iteration time and the accuracy level is found to be about 50-60% improved in recognition compared to the existing neuro classifier.

7. ACKNOWLEDGMENTS

I would like to thanks my guide Prof R.S.Khule, Prof D.D.Dighe (Head of E&TC dept.) and the honorable principal Dr. G.K.Kharate for their valuable time and dedicated support.

Without which it was impossible to complete my paper. Once again I would like to thank you all staff members (E&TC dept.) for their timely support.

8. REFERENCES

[1] R. H. Y. Chung, N. H. C. Yung, and P. Y. S. Cheung, "An efficient parameterless quadrilateral-based image segmentation method," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 9, pp. 1446–1458, Sep.2005.

[2] D. H. Ballard and C. M. Brown, Computer Vision. Englewood Cliffs, NJ: Prentice-Hall, 1982. [3] Bertsekas, D. and Callager.R, (1987) "Data Networks", 5th International conference on computing and communication, pp.325- 333.

[3] A.Bardera, M. Feixas, I. Boada, J. Rigau, and M. Sbert, "Registrationbased segmentation using the information bottleneck method," in Proc. Iberian Conf. Patern Recognition and Image Analysis, June , vol. II, pp.190–197.

[4] P. Bernaola, J. L. Oliver, and R. Román, "Decomposition of DNA sequence complexity," Phys. Rev. Lett., vol. 83, no. 16, pp. 3336–3339,Oct. 1999.

[5] J. Burbea and C. R. Rao, "On the convexity of some divergence measures based on entropy functions," IEEE Trans. Inf. Theory, vol. 28, no.3, pp. 489–495, May 1982.

[6] S. J. Canny, "A computational approach to edge detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 8, no. 6, pp. 679–698, Jun. 1986.

[7] Cocosco, V. Kollokian, R.-S. Kwan, and A. Evans, "Brainweb: Onlineinterface to a 3DMRI simulated brain database," NeuroImage, vol.5, no. 4, 1997.

[8] T. M. Cover and J. A. Thomas, Elements of Information Theory, Wiley Series in Telecommunications. New York: Wiley, 1991

International Journal of Computer Applications Technology and Research
Volume 3– Issue 3, 155 - 158, 2014, ISSN: 2319–8656

The Study of Problems and Applications of Communication Era in (Virtual) E-learning

Amin Ashir
The member of young researchers club, Islamic Azad University of Dezful, Iran

Sedigheh Navaezadeh
Sama Technical and Vocational Training College, Islamic Azad University, Mahshahr, Branch Mahshahr, Iran

Sara Ziagham
Department of Midwifery, Shushtar Faculty of Medical Sciences, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

Abstract: We are in the era called information age. In this era, the role of information and communication is very important because the role of education and training through communication is very effective, and an electronic name has been assigned to the new type of training and learning changes including information gathering, processing and distributing. Interaction of electronic training and knowledge management continuously increases due to unavoidable convergence of these two technologies. In one side, a desired output is the result of learning knowledge integrated with practical skills and experiences. On the other side, if staffs have been trained as well as possible, and be ready for using knowledge, applying and associating it, then knowledge can be managed easily. With regard to benefits of e-learning and its abilities for training, it seems that its integration with current training programs at universities, where common training is provided through integration of traditional learning and e-learning, is unavoidable. This is noticeable in training field that has too many addresses with various interests, experiences and training needs and skills.

Keywords: Information era; E-learning, Knowledge management; Learning knowledge; traditional learning and training

1. INTRODUCTION

E-learning, as an achievement of scholars in science and philosophy fields, is a response to new information requirements, human training in information society. Changing human knowledge paradigm in 20th century and moving from possibility of accessing certain knowledge about the world toward recognition uncertainty is noticeable in technology level. These changes occurring in technology level moved the route of technology development from those technologies that increased human power in mass industrial productions toward those technologies that reinforced thinking power (such as processing, analysis, evaluation and etc.). as an example, e-commerce (in business) and e-learning (in the field of knowledge management) have emerged.

Using information technology in training field requires local standards setting and interdisciplinary e-learning system [1]. Generally, the aim of e-learning is to allow easy, free and searchable access to courses and to improve presentation of course materials and content in order to learn deeply and seriously. Unlike traditional learning and training, in such learning environment, individuals benefit from subjects on the basis of their own abilities. In e-learning, maximum efficiency be obtained by combining and integrating various learning methods such as text, sound, phonemes, picture and etc [2].

Knowledge management and e-learning have common purposes. The aim is to improve individuals learning through training, sharing knowledge, and providing learner organization. Through convergence understanding, many attempts have been done in terms of proper integration. In this paper, we integrate these two fields (knowledge management and e-learning), and express their problems. Knowledge management can create a strong structure and framework for educational content and materials to support e- learning.

2. VIRTUAL LEARNING (E-LEARNING) AND KNOWLEDGE MANAGEMENT

Virtual learning is subset and common subject of information and learning technology. It provides learners continuous learning possibility everywhere and every time. In virtual learning, course presentation and learning is possible through new technologies. According to Davenport theory, knowledge management refers to a systematic process to find, select, organize and represent knowledge in a way that it increases individuals' abilities and capabilities in their area of interest [3].

In knowledge management, organizational view in terms of learning is considered, and it tries to recognize defects in terms of sharing knowledge among individuals of the organization.

3. INTEGRATING VIRTUAL LEARNING AND KNOWLEDGE MANAGEMENT

Knowledge management and e-learning have common purposes. Their aims are to facilitate learning and to provide ability and specialty in an organization. Both

technologies try to present effective knowledge in terms of information and data available in information resources of an organization. In addition, both of them try to improve performance and skills of individuals and groups by distributing knowledge in an organization. Hence, both technologies have a common strategy to create a learner organization. Another common aim of these technologies is the role of interaction, participation and group work of individuals in the organization.

In summary, the role and the effect of knowledge management on e-learning can be explained as follows:

In production cycle and knowledge management of an organization, knowledge can be changed to educational and learning content through using some techniques like grading, catalogue classification, adding the explanation and required interpretations, accommodation with the conditions of knowledge receiver, paying attention to learning and metadata to reuse it. Then, produced educational and learning content is enriched through applying standards, learning and training parameters, motivational parameters and more explanations. Later, a learning scenario is created, and it is presented to individuals, associations and cooperative learning groups. Knowledge presented in the form of learning is integrated with the experience and technical comments of individuals; then, it enters the cycle of knowledge management and e-learning [6].

4. THE PROBLEMS OF INTEGRATING VIRTUAL LEARNING AND KNOWLEDGE MANAGEMENT

Studies and experiences have shown that many ideas presented in terms of integration of knowledge management and e-learning have not been applied and executed due to the following problems and limitations [5]:

- Conceptual level: lack of any conceptual and meaningful relationship between three spaces including work, knowledge and learning.
- Technical level: each above mentioned space involves different traditional and information systems, so integration of these systems is very difficult.
- Ignoring the field: the common problem of knowledge management and e- learning is that, in both of them, the field and conditions in which learning is provided and knowledge is transferred are considered differently. The way and the type of presenting learning and knowledge if different depending on environment conditions, background conditions, preparation, interests, talent and user information.
- Less interaction and cooperation: the problem of applying knowledge management in e- learning is that information parts in the system of knowledge management do not have enough relationship, cohesion, participation and cooperation. It should be considered that conceptual relationship and electronic have great importance in e- learning, and learner participation in learning process to increase learning percent is very important, while information has not been designed on the basis of participatory learning in knowledge management. In order to use information in learning process, learning participatory activities must be considered.
- The problem of dynamic conformity
- Inappropriateness of conceptual and applied content

5. THE PLACE OF ELECTRONIC CONTENT IN E-LEARNING AT UNIVERSITY

Generally, if we consider users (administration agents, teachers, students and supporting), learning processing (and their supporting services) and learning resources as the pillars of e- learning system, then communication and information technologies can be taken into account as an ability maker of this set. In one side, its duty is to provide communication bed and to manage required interactions among these pillars. On the other side, it can be considered as an element to enrich the content. On the basis of executive dimension, the activities of such structures can be divided into learning activities, and training and educational activities (such as administrative activities to support learning. Managing each of them requires special information systems.

With regard to integration approach of research0 learning in learning process at universities, three applications can be considered for learning resources:

1- Facilitation of achievement and reinforcement of information literacy

2- Facilitation of the process of transferring the main concepts and construction of knowledge in considered field.

3- Arranging a condition for being familiar with real world situation in considered field.

In information communities where people need information for their own professional, personal and recreational affairs, one of the main life skills is “information literacy” referring to a set of abilities through which it can be recognized that when and what kind of information is required, and in this way, required information can be evaluated and used effectively [7].

6. VARIOUS APPROACHES OF LEARNING AND TRAINING

There are three main methods of learning and training as follows:

1- Instructional method: in this method, teacher and learning information are emphasized. The aim of this method is to transfer information from teacher to the student. This method is called parrot-like strategy.

2- Constructivist method: in this method, students (learners) are emphasized. Each person makes its own knowledge. In fact, the learner is responsible for its own learning. Teacher plays the role of leader and assistant in learning and training process. This method is called creative thinker strategy.

3- Social constructivist method: in this method, group study in interaction with a community (learning society) is considered with the aim of learning and obtaining knowledge. Learning this method is a process in the form of a social activity.

7. DIFFERENT TYPES OF VIRTUAL LEARNING FOR USERS AND ADDRESSES

Different kinds of this learning are as follows:

* Higher education: the users of this virtual learning are students, teachers, university staff and personnel and even the applicants of higher education courses. Concepts such as virtual and digital university are related to this kind of virtual learning.

* Training aid: the users of this virtual learning are students studying in various educational levels, their parents and teachers. Concepts like virtual schools, virtual high schools and etc. are related to this kind of virtual learning.

* General learning: the users of this virtual learning are ordinary and home users using information technology tools to increase individuals' skills.

* Personnel and staff training and learning: : the users of this virtual learning are personnel and staff of companies, institution, and private and general organizations. Using information technology in the field of education and training of manpower is usually offered to companies, factories and institution where there is much manpower [8].

* Ethics in virtual learning

The aim of ethics in information technology, e- learning and different kinds of it is to provide some tools to use and develop these systems by considering ethical dimensions. Ethics should be defined in the field of psychological knowledge [9] and the science based on respecting the rights of itself and others in interpersonal, intrapersonal and personal interactions [10].

Ethics in e- learning refers to patterns of communication behaviors based on respecting the rights of itself and others. It makes the ethical responsibilities of an organization clear. The rights of others mean internal and external elements of an organization. This organization has interpersonal and intrapersonal interactions. External environment cannot be just reduced to organization customers. Society, government, environment, neighbors and others are beneficiary of the organization [11]. If an organization has interaction in terms of profitability and presenting better services, then it can be considered as external environment. The elements of ethics in virtual learning have been demonstrated in figure 1.

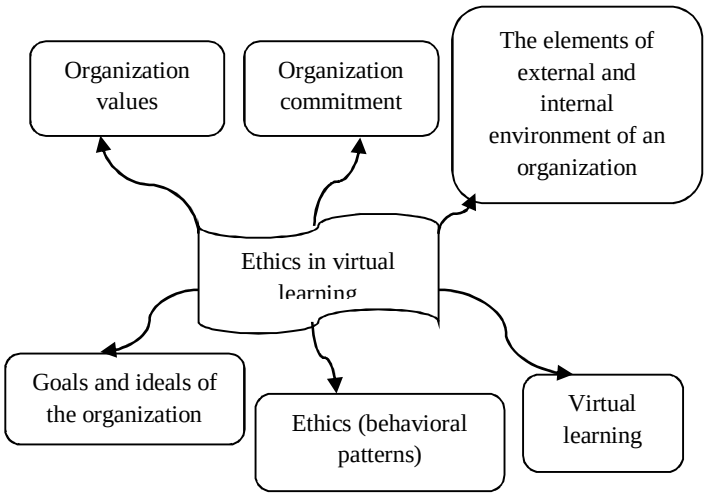

Figure 1: ethical components of virtual learning and training

8. THE COMPONENTS OF VIRTUAL UNIVERSITY

The components of virtual university are as follows:

* Information Booth: it helps students to understand virtual university, its services, its course syllabus and academic and degrees.

* Teaching unit: it refers to offices, and training and educational units presenting courses, seminars, laboratories, thesis and examination programs.

* Students' office: it is responsible of administrative and executive services such as recording courses, seminars, examinations and workshops.

* Digital library: through digital library, information lists of library can be accessed.

* Cafeteria: it provides students communication as well as discussion.

* Blackboard: students can be informed of news.

* Research center: this center informs students about research activities and publications. Also, it provides communication between students and researchers.

* Shop: this place facilitates buying course resources.

9. SOME CHARACTERISTICS OF VIRTUAL UNIVERSITY

Virtual university refers to an environment presenting e-learning services through using appropriate multimedia tools and communication structure. Some characteristics of virtual university are as follows:

* There is no need to physical presence of teacher and student in class.
* The higher quality of course syllabus
* Supporting many students in a class
* It is economical, and access is easy.

10. CONCLUSION AND SUGGESTIONS

Understanding e-learning and knowledge management on the basis of their definitions is easy. These technologies can able the organization to manage their own knowledge capitals in production cycle of knowledge, training and learning content, and to transfer and share the content. Presenting training on the basis of business and knowledge requirements of staff and consistent with their interests and priorities has been considered by the organization in terms of integration of two concepts including e- learning and knowledge management. By integrating these technologies, learner organization using its own knowledge assets can be created. With regard to characteristics and capabilities of virtual environment and the role of virtual teacher in this environment, teacher encourages the students to cooperative learning. S/he participates in discussions as a mediator, and initiates discussion when necessary. Virtual teacher should design various learning activities, and should introduce reliable and valid resources to help and encourage students to participate actively in learning. Therefore, teaching strategies in course syllabus of virtual university must be selected according to the following instructions:

* They must increase the interaction between the teacher and student as well as their cooperation with each other.
* They must motivate the students to learn actively.
* They allow the teacher to pay attention to students by quick reaction.
* Individual differences of students must be considered.
* Cognitive flexibility should be reinforced in students.
* They must be selected on the basis of problem-oriented methods and emphasizing on learning methods.
* They should facilitate the interaction between the learner and various resources of learning.
* Generally, the aim of training is to propagate ethics, but free communication in virtual learning and training, and emergence of unethical behaviors require paying attention to ethics. In discussion of ethics in virtual training, complete and exact conceptualization of this word is required. In conceptualization of ethics in virtual training, some factors have important place such as paying attention to ethics, virtual learning and training, the elements of internal and external organization environment, values, commitment and organizational goals and ideals because paying attention to just one dimension causes transition and change.

11. REFERENCES

[1] Standard institution and industrial researches of Iran, E-learning (Virtual Learning)- Characteristics. Tehran, 2010.

[2] Parinaz bani Si, Seddighe Mollaeian, Fatemeh Peikarifar. The first student conference of e- learning, science and industry university

[3] Okamoto Toshio, Ninomiya Toshie- Organization knowledge management system for e-learning practicein universities-IEEE Paper-Proceedings the sixth conference IASTED Interactional. Conference Web-Based Education, chamonix.France Year of Publication:2007- Volume2 -PP. 528- 536.

[4] Stefanie N. Lindstaedito.Johannes Farmer-integration Knowledge management and e-Learning. UCSS SpeciaI ls sue-Journal Universal Computer Science vol. 11. no.3 (2005)375-377submitted:3/3/05, accepted:17/3/05 appeared: 28/3/05 J .UCS-pp.375

[5] Ras Eric, Memmel Martin.Weibelzahl Stephan(Eds.)-integration of E-Learning and Knowledge Management - Barriers. Solution and Future lssues- Vol 3782/2005 -A thoff et al.(Eds.)WM 2005.L NAI 3782,2 005. Springer-Verlag Berlin Heidelberg 2005-K. D - pp. 155- 164.

[6] Miltiades D. Lytras, Ambjorn Naeve, Athanasius Pouloudi-Knowledge Management as a Reference Theory for ELearning:A Conceptual and Technological Perspective-Interactional Journal of Distance Education Technologies.3(2),l -12,April- June2 005.Copyright@2005-pp.1-12.

[7] American Library Association. Presidential Committee on Information Literacy. Final Report.(Chicago: American Library Association, 1989.) 1 Information.

[8]Robabeh Farhady. E-learning as a New Paradigm in Communication Era. Periodical of science and information technology, pages 49-66, 2006.

[9] Nima Ghrbani. Communication Styles and Skills. Tehran: tenth punlication of Tabalvor, 2006.

[10] Faramarz Ghara Maleki. Professiobnal Ethics. Tehran, 2004

[11] Mohammad Mehr Mohammadi. Course syllabus, Perspectives and approaches. Tehran: Astane Ghodse Razavi, 2005.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 3, 159 - 164, 2014, ISSN: 2319–8656

Adaptive Neural Fuzzy Inference System for Employability Assessment

Rajani Kumari
St Xavier's College,
Jaipur, India

Vivek Kumar Sharma
Jagannath University,
Jaipur, India

Sandeep Kumar
Jagannath University
Jaipur, India

Abstract: Employability is potential of a person for gaining and maintains employment. Employability is measure through the education, personal development and understanding power. Employability is not the similar as ahead a graduate job, moderately it implies something almost the capacity of the graduate to function in an employment and be capable to move between jobs, therefore remaining employable through their life. This paper introduced a new adaptive neural fuzzy inference system for assessment of employability with the help of some neuro fuzzy rules. The purpose and scope of this research is to examine the level of employability. The concern research use both fuzzy inference systems and artificial neural network which is known as neuro fuzzy technique for solve the problem of employability assessment. This paper use three employability skills as input and find a crisp value as output which indicates the glassy of employee. It uses twenty seven neuro fuzzy rules, with the help of Sugeno type inference in Mat-lab and finds single value output. The proposed system is named as Adaptive Neural Fuzzy Inference System for Employability Assessment (ANFISEA).

Keywords: Neural Network, Fuzzy Logic, Employability, Sugeno type inference, Education, Understanding Power, Personal Development

1. INTRODUCTION

The problem of finding membership functions and fitting rules is frequently a demanding process of endeavor and error. This leads to the idea of applying knowledge algorithms to the fuzzy systems. The neural networks which have efficient learning algorithms had been obtainable as an alternative to computerize or to maintain the development of tuning fuzzy systems. Progressively, its application extent for all the areas of the knowledge in the vein of data classification, data analysis, imperfections detection and maintain to decision-making. JSR Jang proposed an adaptive network based fuzzy inference system [4]. The architecture and knowledge procedure underlying ANFIS is offered, which is a fuzzy inference system employed in the framework of adaptive networks. By by means of a hybrid knowledge procedure, the proposed ANFIS can build an input-output plotting based on both human knowledge and specified input-output data pairs. CF Juang proposed an online self-constructing neural fuzzy inference network and its applications [7]. It proposed a self-constructing neural fuzzy inference network (SONFIN) through online knowledge ability. The SONFIN is naturally a modified Takagi Sugeno Kang type fuzzy rule based model holding neural network knowledge ability. NK Kasabov, and Q Song proposed a dynamic progressing neural fuzzy inference system and its application for time series prediction [9]. It introduces an innovative type of fuzzy inference systems which indicated as dynamic evolving neural fuzzy inference system (DENFIS) for adaptive online and offline knowledge and their application for dynamic time series forecast. CT Lin and CS Lee proposed a neural network based fuzzy logic control and decision system [10]. This model associate the notion of fuzzy logic controller and neural network configuration in the form of feed forward multilayer net and knowledge abilities into an incorporated neural network based fuzzy logic control and decision system. O Avatefipour et al. designed a New Robust Self Tuning Fuzzy Backstopping Methodology [11]. It is focused on suggested Proportional Integral (PI) like fuzzy adaptive backstopping fuzzy algorithm constructed on Proportional Derivative (PD) fuzzy rule base through the adaptation laws consequent in the lyapunov sense. GO Tirian et al. proposed an adaptive control system for continuous steel casting based on neural networks and fuzzy logic [12]. It defines a neural network based approach for crack extrapolation aimed at improving the steel casting process presentation by decreasing the number of crack produced by failure cases. A neural system to approximation crack detection possibility has been designed, implemented, tested and incorporated into an adaptive control system. R Kumari et al. applied fuzzy control system for scheduling CPU [41], Job Shop scheduling [42], two way ducting system [40] and air conditioning system [43].

2. EMPLOYABILITY

Employability is defined as a set of accomplishments which consider skills, understandings and personal attributes. These achievements are make graduates further likely to gain employment and be prosperous in their selected occupations. Employability skills are generic or non-technical skills, such as communication, team work, self-management, planning and organizing, positive attitude, learning, numeracy, information technology and problem solving, which subsidize to your ability to be a successful and effective participant in the workplace. They are occasionally referred to as key, core, life, essential, or soft skills. Many employability skills and technical skills are exchangeable between jobs. Employability plays a significant role in the implementation of the Teaching Strategies and College Learning. It is part of worthy learning exercise. Students who involve in emerging their employability are likely to be reflective, independent and responsible learners. Teaching, innovative learning and assessment approaches which encourage students'

understanding and help them to participate in deep learning will also improve their employability. Concerning employers in the education knowledge can help students appreciate the significance of their course and acquire how to apply knowledge and theory in practical ways in the workplace. R Kumari et al. proposed an expert system for employability [45] and a fuzzified employability assessment system [44].

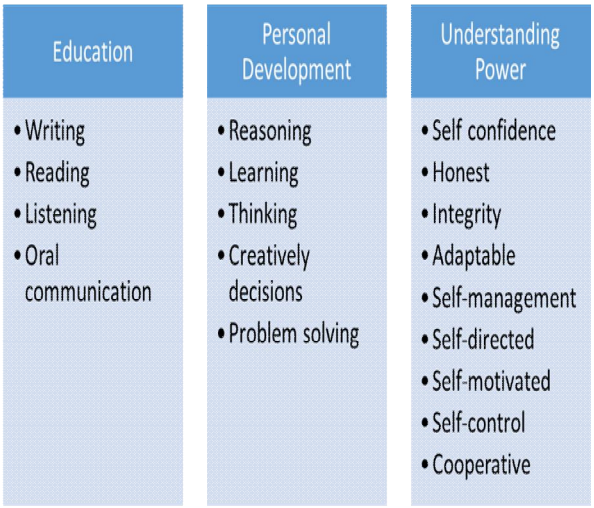

Figure 1. Classification of Employability Skills

3. FUZZY LOGIC CONTROL SYSTEM

Fuzzy systems recommend a mathematic calculus to interpret the subjective human awareness of the real processes. This is a way to control the practical awareness with some level of improbability. The Fuzzy logic techniques were firstly recommended by A L Zadeh in 1956 [1] [2] [3]. Aim of these techniques were scheming a system in which employers are permitted to form sets of rules through linguistic variables and membership functions, after that, the system renovates these rules into their mathematical complements.

4. NEURO FUZZY LOGIC CONTROL SYSTEM

Fuzzy logic and artificial neural networks [5][6] both are analogous tools for crafting systems that deal with expectation and classification of tasks. The idea of different terminologies for neuro-fuzzy systems introduced in the literature was neuro-fuzzy systems [8]. The term neuro-fuzzy system is usually a shortening of adaptive fuzzy systems industrialized by manipulating the similarities among fuzzy systems and neural networks methods. The two techniques of fuzzy logic and neural networks have combined in several different ways. In general, there are three combinations of these techniques. One is neural-fuzzy systems, another one is fuzzy neural networks and third one is fuzzy-neural hybrid systems. Neuro-fuzzy architecture Fuzzy Adaptive Learning Control Network (FALCON) proposed by CT Lin and CS Lee [30]. Architecture Adaptive Network based Fuzzy Inference System (ANFIS) proposed by R. R. Jang [31]. Architecture Neuronal Fuzzy Controller (NEFCON) proposed by D. Nauck and Kruse [32]. Architecture Fuzzy Net (FUN) proposed by S. Sulzberger, N. Tschichold and S. Vestli [33]. Architecture Fuzzy Inference and Neural Network in Fuzzy Inference Software (FINEST) proposed by O Tano and Arnould [34]. Architecture of Self Constructing Neural Fuzzy Inference Network (SONFIN) proposed by Juang and Lin [35]. Architecture Dynamic/Evolving Fuzzy Neural Network (EFuNN and dmEFuNN) proposed by Kasabov and Song [36]. Architecture Generalized Approximate Reasoning based Intelligence Control (GARIC) proposed by H. Berenji [29]. Architecture Fuzzy Neural Network (NFN) proposed by Figueiredo and Gomide [37].

4.1 Neural Fuzzy System

The neural network is used to regulate the functions and representing the fuzzy sets which are operated as fuzzy rules. The neural network deviation its weight in the training for the expectation of diminishing the mean square error amid the tangible output of the networks and the targets. L. Wang, J. Mendel, Y. Shi and M. Mizumoto proposed some illustrations of this approach [19, 20, 21]. Neural fuzzy systems are used in controller systems.

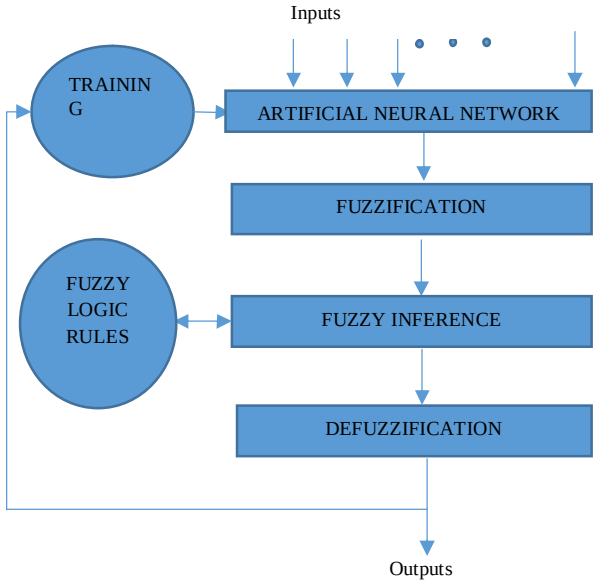

Figure 2. Neural Fuzzy System

4.2 Fuzzy Neural Network

A fuzzy neural network introduced memory connections for classification and weight connections for selection, so that it solves concurrently two foremost problems in pattern recognition that is pattern classification and feature selection. Fuzzy neural systems are used in pattern recognition applications. Lin and Lee presented a neural network in 1996 which composed of fuzzy neurons [16].

4.3 Fuzzy Neural Hybrid System

A fuzzy neural hybrid system is prepared individually from both fuzzy logic and neural network techniques to bring out solicitations such as control systems and pattern recognition. The lead objective of the fuzzy neural hybrid system can be proficient by having each technique do its task by incorporating and approving one another. This kind of inclusion is application oriented and appropriate for control and pattern recognition applications both. The worthy example of hybrid neuro fuzzy are GARIC, ARIC, ANFIS the NNDFR model [22, 23, 18, 38, 17].

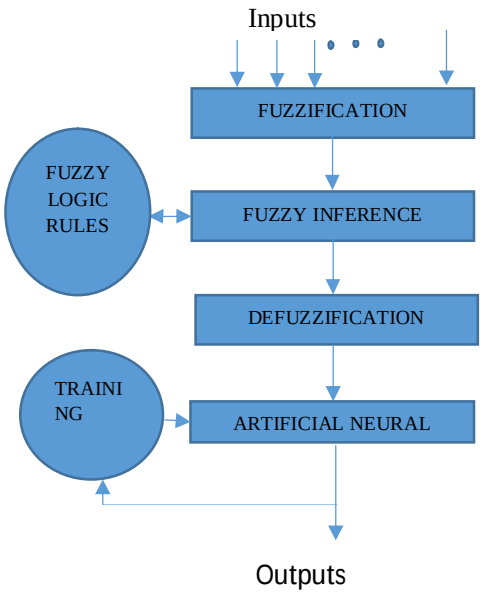

Figure 3. Fuzzy Neural System

5. ANFIS STRUCTURE

The adaptive neuro fuzzy inference system (ANFIS) is a commercial approach which is combined the two techniques such as a neural network and a fuzzy logic to generate a complete shell [18] Fundamentally the system of ANFIS applies the method of the artificial neural network learning rules to conclude and adjust the fuzzy inference systems parameters and structure. Many important features of ANFIS can support the system to achieve a task intensely; these features are considered as fast and accurate learning, easy to implement, excellent explanation facilities, strong generalization abilities, through fuzzy rules. It is easy to integrate both linguistic and numeric acquaintance for problem solving [18, 38, 39, 13, 14, 15]. This system is measured as an adaptive fuzzy inference system through the competency of learning fuzzy rules from data and as a connectionist manner provided with linguistic significance. A hybrid neuro-fuzzy inference expert system had developed by Jang that works in Takagi-Sugeno type fuzzy inference system [24, 25, 26, 27, 28]. ANFIS method is used as a teaching technique for Sugeno-type fuzzy systems. System constraints are identified by the support of ANFIS. When ANFIS is applying, generally the number and type of fuzzy system membership functions are well defined by user. ANFIS technique is a hybrid technique, which consists two parts, one is gradient technique which is applied to calculation of input membership function parameters, and another one is least square technique which is applied to calculation of output function parameters.

6. FUZZIFIED EXPERT SYSTEM FOR EMPLOYABILITY ASSESSMENT

In the previous research work initiates a new expert system for assessment of employability with the help of some fuzzy rules. These rules are ultimately used for observe the optimal valuation for employability. This employability compacts with various fuzzy rules and these rules are constructed on employability skills. It computes the Employability Skills for several employees with the help of Mamdani type inference. It used linguistic variables as input and output for calculate a crisp value for employability skills.

7. ADAPTIVE NEURAL FUZZY INFERENCE SYSTEM FOR EMPLOYABILITY ASSESSMENT

This paper introduced an innovative adaptive neural fuzzy inference system for employability with the help of some neuro fuzzy rules. These neuro fuzzy rules are ultimately used for examine the best valuation for employability. This employability deals with some neuro fuzzy rules and these rules are based on three employability skills named as education, Personal Development and Understanding Power. This work is proposed to compute the Employability Level for any employee with the help of Takagi Sugeno type inference.

This concern research use suitable linguistic variables as input and output for calculate a crisp value for employability. Education (E), Personal Development (PD) and Understanding Power (UP) measured as Low, Medium and High and Employability skills (ES) measured as Very Low, Low, Medium, High and Very High. The recommended skills is a gathering of linguistic neuro fuzzy rules which designate the relationship between distinct input variables (E, PD and UP) and output (ES).

Table 1. Membership function and range of input variables

Education	Personal Development	Understanding Power	Range
Low	Low	Low	0-4
Medium	Medium	Medium	2-8
High	High	High	6-10

Table 2. Membership function and range of output variable

Employability	Range
Very Low	0-2
Low	1-4
Medium	3-6
High	5-8
Very High	7-10

Table 1 encloses the membership functions and range of input variables named as education, employability and understanding power. Table 2 encloses membership function and range of output variable named as employability. Table 3 encloses the twenty seven rules which are built on IF THEN statement such as

IF E is high and PD is high and UP is high THEN ES is high

These rules are used for calculate the crisp value using centroid defuzzification technique of Sugeno type inference in Matlab that signifies the employability level of each and every employee.

Figure 4 shows the membership function of input variable education, figure 5 shows input variable personal development, figure 6 shows input variable understanding power, figure 7 shows output variable employability, figure 8

shows ANFIS structure and figure 9 outlines rules of employability.

Table 3. Set of proposed rules

Rule Number	Education	Personal Development	Understanding Power	Employability
1	Low	Low	Low	Very Low
2	Low	Low	Medium	Very Low
3	Low	Low	High	Low
4	Low	Medium	Low	Very Low
5	Low	Medium	Medium	Medium
6	Low	Medium	High	Medium
7	Low	High	Low	Low
8	Low	High	Medium	Medium
9	Low	High	High	Medium
10	Medium	Low	Low	Very Low
11	Medium	Low	Medium	Low
12	Medium	Low	High	Medium
13	Medium	Medium	Low	Medium
14	Medium	Medium	Medium	Medium
15	Medium	Medium	High	High
16	Medium	High	Low	Medium
17	Medium	High	Medium	High
18	Medium	High	High	Very High
19	High	Low	Low	Low
20	High	Low	Medium	Medium
21	High	Low	High	Medium
22	High	Medium	Low	High
23	High	Medium	Medium	High
24	High	Medium	High	Very High
25	High	High	Low	High
26	High	High	Medium	Very High
27	High	High	High	Very High

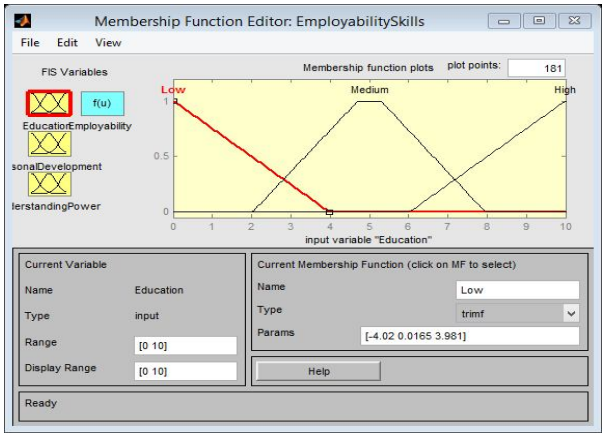

Figure 4. Input Variable "Education"

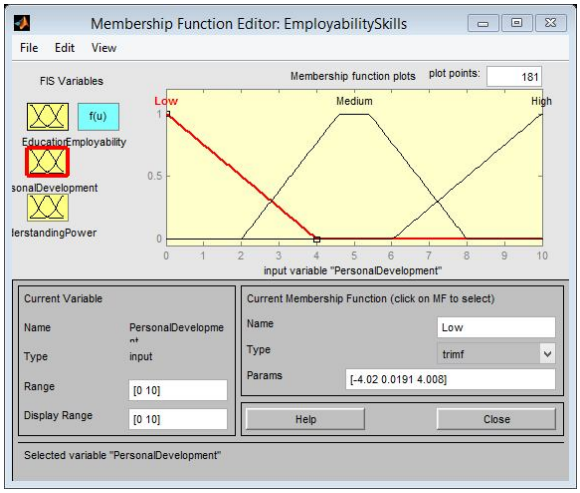

Figure 5. Input Variable "Personal Development"

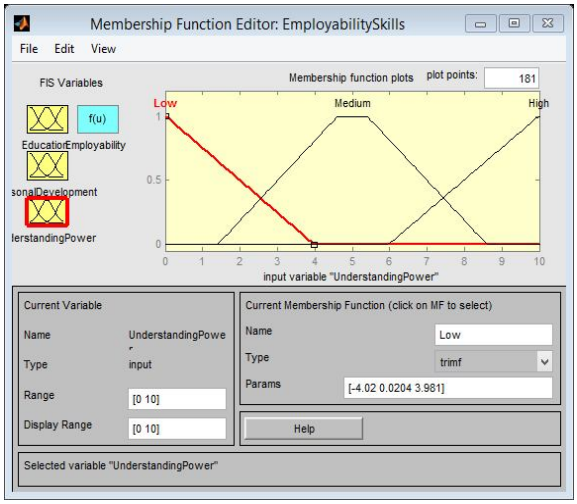

Figure 6. Input Variable "Understanding Power"

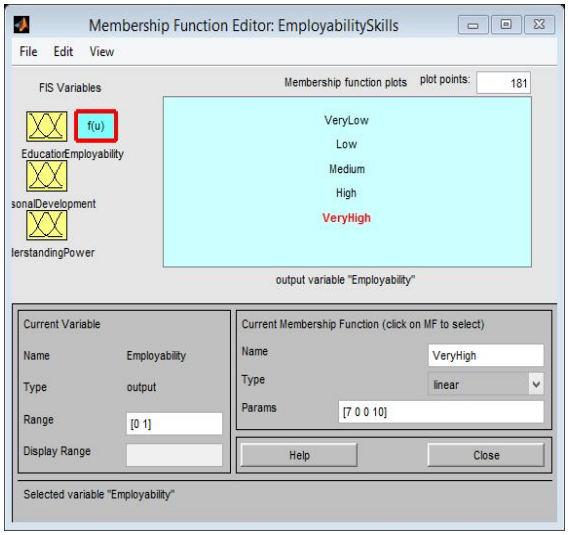

Figure 7. Output Variable "Employability"

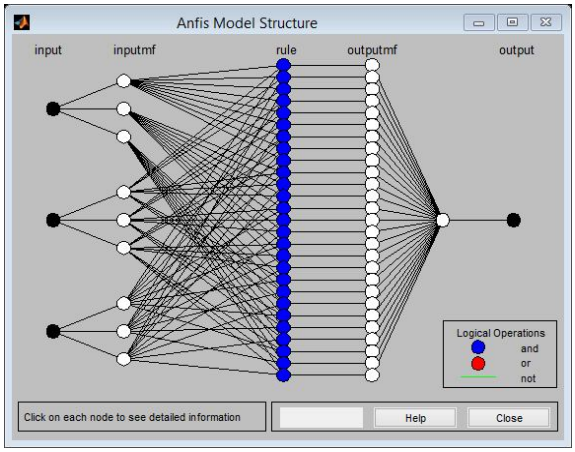

Figure 8. ANFIS Structure for Employability

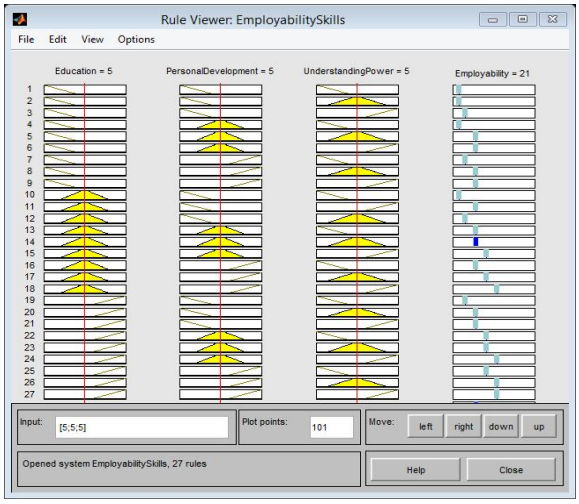

Figure 9. Rules for Employability Skills

8. CONCULSION

This paper estimated an adaptive neural fuzzy inference system for employability assessment. The concern research finds the level or capability of any employee with the help of three employability skills named as education, personal development and understanding power. The proposed system is beneficial for organization to compute employability level for individual in a simple manner. With the help of proposed system employer can simply filter best appropriate candidates based on their education, personal development and understanding power. This system operates above three inputs based on neuro fuzzy rules and computes employability.

9. REFERENCES

[1] Zadeh, L.. Fuzzy Sets. Inf Cont, Vol. 8, Pp. 338–353, 1965.

[2] Royas, I.; Pomares, H.; Ortega, J.; And Prieto, A. (2000). Self-Organized Fuzzy System Generation From Training Examples, Ieee Trans. On Fuzzy Systems, Vol. 8, No. 1, Pp. 23-36, 2000.

[3] Cox, E. , The Fuzzy Systems Handbook. Ap Professional - New York.1994.

[4] Jang, J-SR. "ANFIS: adaptive-network-based fuzzy inference system."*Systems, Man and Cybernetics, IEEE Transactions on* 23.3 (1993): 665-685.

[5] Haykin, S. , Neural Networks, A Comprehensive Foudation. Second Edition, Prentice Hall. 1998.

[6] Mehrotra, K., Mohan, C. K., And Ranka, S. ,Elements Of Artificial Neural Networks. The Mit Press, 1997

[7] Juang, Chia-Feng, and Chin-Teng Lin. "An online self-constructing neural fuzzy inference network and its applications." *Fuzzy Systems, IEEE Transactions on*6.1 (1998): 12-32.

[8] Buckley, J.J. & Eslami, E., Fuzzy Neural Networks: Capabilities. In Fuzzy Modeliparadigms And Practice , Pedrycz W, Ed., Pp. 167-183, Kluwer, Boston, 1996.

[9] Kasabov, Nikola K., and Qun Song. "DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction." *Fuzzy Systems, IEEE Transactions on* 10.2 (2002): 144-154.

[10] Lin, C-T., and C. S. George Lee. "Neural-network-based fuzzy logic control and decision system." *Computers, IEEE Transactions on* 40.12 (1991): 1320-1336.

[11] Avatefipour, Omid, et al. "Design New Robust Self Tuning Fuzzy Backstopping Methodology." (2014).

[12] Tirian, Gelu-Ovidiu, Ioan Filip, and Gabriela Proştean. "Adaptive control system for continuous steel casting based on neural networks and fuzzy logic."*Neurocomputing* 125 (2014): 236-245.

[13] Jang, J.S.R; Sun, C.T & Mizutani, E. , Neuro-Fuzzy And Soft Computin. Prentice-Hall: Englewood Cliffs, Nj, 1997.

[14] Lin, C.T. & Lee, C.S., Neural-Network-Based Fuzzy Logic Control And Decision Systems. Ieee Trans. On Computers, Vol. 40, No. 12, Pp. 1320-1336, 1991

[15] Lin, C.T. And Lee, G., Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent systems. Ed. Prentice Hall, 1996.

[16] Lin, C.T. And Lee, G., Neural Fuzzy Systems: A Neuro-Fuzzy Synergism To Intelligent Systems. Ed. Prentice Hall. 1996.

[17] Takagi, H. & Hayashi, I., Nn-Driven Fuzzy Reasoning. International Journal Of Approximate Reasoning, Vol. 5, Issue 3, 1991.

[18] Jang, J.S.R. & Sun, C.T., Functional Equivalence Between Radial Basis Function Networks And Fuzzy Inference Systems, Ieee Trans. On Neural Networks, Vol. 4, No. 1, Pp. 156-159, 1993.

[19] Wang, L. And Mendel, J., Back-Propagation Fuzzy System As Nonlinear Dynamic System Identifiers. Proceedings Of Ieee International Conferenceon Fuzzy Systems, Pages 1409–1416, 1992

[20] Shi, Y. And Mizumoto, M. (2000a). A New Approach Of Neurofuzzy Learning Algorithm For Tuning Fuzzy Rules. Fuzzy Sets And Systems, 112(1):99–116, 2000a.

[21] Shi, Y. And Mizumoto, M., Some Considerations On Conventional Neuro-Fuzzy Learning Algorithms By Gradient Descent Method. Fuzzy Sets And Systems, Vol. 112, No. 1, Pp. 51–63, 2000b.

[22] Berenji, R.H. , A Reinforcement Learning-Based Architecture For Fuzzy Logic Control. International Journal Of Approximate Reasoning, Vol. 6, Issue 2, 1992.

[23] Bersini H.; Nordvik, J.P & Bonarini, A. , A Simple Direct Adaptive Fuzzy Controller Derived From Its Neutral Equivalent, Proceedings Of 2 Ieee International Conference On Fuzzy Systems, Vol. 1, Pp. 345-350. Nd, 1993.

[24] Abraham A., "Adaptation Of Fuzzy Inference System Using Neural Learning, Fuzzy System Engineering: Theory And Practice", Nadia Nedjah Et Al. (Eds.), Studies In Fuzziness And Soft Computing, Springer Verlag Germany, Isbn 3-540-25322-X, Chapter 3,Pp. 53–83, 2005.

[25] Tharwat E. Alhanafy, Fareed Zaghlool And Abdou Saad El Din Moustafa, Neuro Fuzzy Modeling Scheme For The Prediction Of Air Pollution, Journal Of American Science, 6(12) 2010.

[26] T. M. Nazmy, H. El-Messiry, B. Al-Bokhity, Adaptive Neuro-Fuzzy Inference System For Classification Of Ecg Signals, Journal Of Theoretical And Applied Information Technology, Pp-71-76, 2010.

[27] Abdulkadir Sengur., "An Expert System Based On Linear Discriminant Analysis And Adaptive Neurofuzzy Inference System To Diagnosis Heart Valve Diseases, Expert Systems With Applications, 2008.

[28] G. Zhao, C. Peng And Xiting Wang., "Intelligent Control For Amt Based On Driver's Intention And Anfis Decision-Making," World Congress On Intelligent Control And Automation, 2008.

[29] H. R. Berenji and P. Khedkar, "Learning and Tuning Fuzzy Logic Controllers through Reinforcements", IEEE Transactions on Neural Networks, 1992, Vol. 3, pp. 724-740.

[30] T. C. Lin, C. S. Lee, "Neural Network Based Fuzzy Logic Control and Decision System",IEEE Transactions on Computers, 1991, Vol. 40, no. 12, pp. 1320-1336.

[31] R. Jang, "Neuro-Fuzzy Modelling: Architectures, Analysis and Applications", PhD Thesis, University of California, Berkley, July 1992.

[32] D. Nauck, R, Kurse, "Neuro-Fuzzy Systems for Function Approximation", 4th International Workshop Fuzzy-Neuro Systems, 1997.

[33] S. Sulzberger, N. Tschichold e S. Vestli, "FUN: Optimization of Fuzzy Rule Based Systems Using Neural Networks", Proceedings of IEEE Conference on Neural Networks, San Francisco, March 1993, pp. 312-316.

[34] S. Tano, T. Oyama, T. Arnould, "Deep Combination of Fuzzy Inference and Neural Network in Fuzzy Inference", Fuzzy Sets and Systems, 1996, Vol. 82(2), pp. 151-160.

[35] F. C. Juang, T. Chin Lin, "An On-Line Self Constructing Neural Fuzzy Inference Network and its applications", IEEE Transactions on Fuzzy Systems, 1998, Vol. 6, pp. 12-32.

[36] N. Kasabov e Qun Song, "Dynamic Evolving Fuzzy Neural Networks with 'm-out-of-n' Activation Nodes for On-Line Adaptive Systems", Technical Report TR99/04, Departement of Information Science, University of Otago, 1999.

[37] M. Figueiredo and F. Gomide; "Design of Fuzzy Systems Using Neuro-Fuzzy Networks", IEEE Transactions on Neural Networks, 1999, Vol. 10, no. 4, pp.815-827.

[38] Jang, J.S.R., Anfis: Adaptive-Network-Based Fuzzy Inference System, Ieee Transactions On Systems, Man And Cybernetics, Vol. 23, No.3, Pp. 665–685. 1993.

[39] Jang, J.S.R. & Sun, C.T., Neuro-Fuzzy Modeling and Control, Proceedings Of The Ieee, Vol. 83, Pp. 378-406, 1995.

[40] R Kumari, S Kumar and VK Sharma. "Two Way Ducting System Using Fuzzy Logic Control System." *international journal of electronics* (2013).

[41] R Kumari, VK Sharma, and S Kumar. "Design and Implementation of Modified Fuzzy based CPU Scheduling Algorithm."International Journal of Computer Applications 77.17 (2013): 1-6.

[42] R Kumari, VK Sharma, S Kumar, Fuzzified Job Shop Scheduling Algorithm, HCTL Open International Journal of Technology Innovations and Research, Volume 7, January 2014, ISSN: 2321-1814, ISBN: 978-1-62951-250-1.

[43] Rajani Kumari, Sandeep Kumar, Vivek Kumar Sharma: Air Conditioning System with Fuzzy Logic and Neuro-Fuzzy Algorithm. SocProS 2012: 233-242

[44] R Kumari. VK Sharma, S Kumar, Employability Valuation Through Fuzzification, in Proceeding of National Conference on Contextual Education and Employability, February 11-12, 2014.

[45] R Kumari. VK Sharma, S Kumar, Fuzzified Expert System for Employability Assessment. Unpublished.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 3, 165 - 168, 2014, ISSN: 2319–8656

Re-enactment of Newspaper Articles

Thilagavathi .N
Sri ManakulaVinayagar
Engineering College
Pudhucherry, India

Archanaa S.R
Sri ManakulaVinayagar
Engineering College
Pudhucherry, India

Lavanya.K
Sri ManakulaVinayagar
Engineering College
Pudhucherry, India

Valarmathi.S
Sri ManakulaVinayagar
Engineering College
Pudhucherry, India

Abstract: Every document that we use has become digitized which makes a great way to save, retrieve and protect documents. They are digitized to have a backup for most paper work .Digitization is found to be more important since everything is going paper free. Digitization of newspaper contributes greatly to preservation and access to newspaper archives. Our paper provides an integrated mechanism that involves document image analysis and k means clustering algorithm to digitize news articles and provide an efficient retrieval of user requested news article. In first stage the news article is segmented from newspaper and pre-processed. In the second stage the pre-processed news articles are clustered by K-means clustering algorithm and key words are extracted for each cluster. The third stage involves selection of cluster containing key phrase given by user and providing the user with requested news article.

Keywords: Page segmentation, TF-IDF weighting, Cosine similarity, Clustering, K-Means algorithm, Keyword Extraction.

1. INTRODUCTION

Document digitization plays a vital role in electronic publishing of newspaper. Digitization of newspaper has become very essential to protect historical news articles, easy storage and efficient retrieval of news articles when needed. In order to obtain the above functionalities, the digitized newspaper need to be powered up with algorithms for document image analysis , efficient storage and retrieval to avoid the ambiguousness during the retrieval of specific news article. Moreover transferring the news article into the system by hand consumes more time and human resource. Thus there is a need for an automated system to obtain the above functionalities.

The basic unit of newspaper is composed of news articles. Document Image Analysis is done to obtain the articles from each and every section of the newspaper one by one. This task is very challenging since it needs to consider the syntactic and semantic information of the blocks of content in every news article. Using the syntactic and semantic information from the image analysis, the newspaper is segmented into individual news article. The content of the each segmented news article is converted into a word file and stored into the database using clustering algorithm. Clustering of news articles involves grouping of news articles into clusters, where they share common properties and keywords. Here, we implement k-means clustering algorithm which is suitable for huge data set like digitized newspapers of years and years. Further, the keyword for each cluster is determined for the efficient retrieval of the required article based on search phrase provided by the user.

2. RELATED WORKS

There are many researches done on newspaper digitization, storage of digitized newspaper and efficient retrieval of them. Most of the existing system does not combine best approach for all three processes together. LiangcaiGao et al. proposed a method to reconstruct Chinese newspaper [5] by accomplishing several tasks such a article body grouping, reading order detection, title-body association and linking scattered article by travelling salesman problem (TSP) and Max-Min Ant System (MMAS). In order to increase the efficiency of MMAS a level based pheromone mechanism is done. It includes two subtask enactment of news article in reverse order by detecting reading order and then using the content continuity to aggregate the text blocks. This method is time prone since it involves semantic analysis of the newspaper content to separate the news article from newspaper. Fu Chang et al. established an approach for layout analysis using adaptive regrouping strategy for Chinese document [2]. This method is specific for Chinese documents that involve horizontal as well as vertical text lines. Wei-Yuan Chen uses an adaptive segmentation method to extract text blocks from colored journals [1] which involves RLSA (run-length smoothing algorithm).This approach needs improvement to adjust the segmentation of non-uniformly colored character from background with complex color.

Osama Abu Abbas proposed a comparison between the four major clustering algorithm k-means algorithm, hierarchical clustering algorithm, self-organization map (SOM) algorithm and expectation maximization algorithm [3]. These algorithms were selected for comparison based on their popularity, flexibility, applicability and handling high dimensionality. These algorithms was compared based on size of dataset, number of clusters, type of dataset the algorithms are going to handle and type of software those algorithm is to be implemented. The result shows that the k-means clustering algorithm is known to be efficient in clustering large data sets. The k-means algorithm allows discovery of clusters from subspaces by identifying the weights of its variables and it is also efficient in identifying noise variables in data. K-means algorithm is suitable for variable selection in data mining. FarzadFarahmandnia proposes a method for automatic key word extraction in text mining using WordNet[4]. By this method the text files are normalized by TDIDF algorithm and preprocessed to remove stop words. Then each word in the text file are hierarchically structured in WordNet dictionary .In order to avoid ambiguities between search words in hypernym search, comparison of every pair of words in document is done. This is done by determining the distance between the two words which is calculated by number of edges between node nodes with search word. Thus words with much closer distance will be chosen as key words for the text document. This paper proposes an approach to segmentation of news article from newspaper, clustering of news article based on its content and assigning labels for each cluster using WordNet.

3. SYSTEM ARCHITECTURE

The proposed system uses scanned newspaper images as input. A newspaper page image contains many articles. These articles are segmented from newspaper using the method, article segmentation by which each news article is made as a text file. These text file is preprocessed to remove stop words and stem words. In order to compare the text documents to compute similarity we perform TF-IDF weighting and cosine similarity. Based on the similarity between the documents, K-means algorithm is used. It is done to cluster documents that express maximum similarity. Keywords for each cluster are extracted to enhance searching. When the user query for a news article the requested news article is retrieved based on key word matching, post processing method involving WordNet.

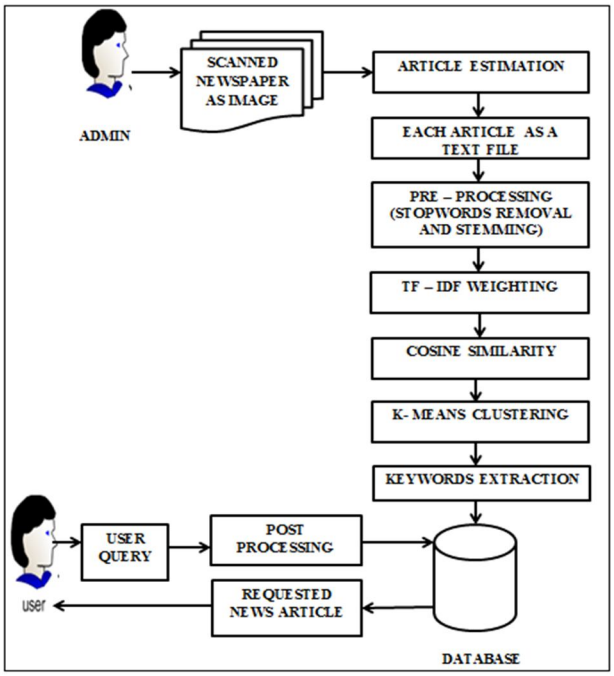

Figure 1. Architecture diagram of Re-Enactment of newspaper article

4. RESEARCH PROPOSAL

4.1 Page Segmentation

To start page segmentation to obtain news article from scanned newspaper image the first essential element to be identified are horizontal and vertical foreground lines. They indicated the boundary of the news article in a newspaper. In order to identify the boundary, binary image of newspaper is transferred into grayscale image. The grayscale image is sub-sampled with respect to foreground pixel. From the result, we obtain two images by assigning all foreground pixel with the length of vertical or length of horizontal line. Thus the horizontal and vertical line needs to be identified are resulted. It is identified since the sub-sampled gray scale image is applied with a condition that is to obtain only pixels whose length or width is larger or smaller respectively than the threshold. Thus the pixel featuring only the horizontal and vertical boundaries of the article is obtained as result. The final stage of segmentation is to extract text from the segmented image. The result of sub-sampling with respect to background pixels are used in order to avoid extracting text from neighbor block. Each block of image is given as an input to OCR (Optical Character Recognizer) which converts each article into a text document.

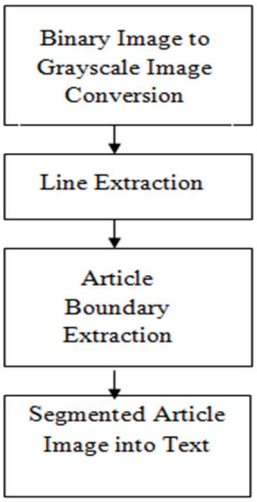

Figure 2. Segmentation of news article from newspaper

4.2 Pre-Processing Of Documents

For every article in the scanned newspaper, a word document is created. Pre-processing of these word file has to be done to prune words from the document with poor information. It optimizes the keyword list that contain list of terms in the document. It involves removal of stop words and stemming words. Pronouns, preposition, conjunction and punctuations carry no meaning as keywords are to be removed in pre-processing. The words in the document are listed out and if it is present in the list of stop words that has been pre-defined in our method, they are removed. This is followed by removal of stemming words. It involves finding variant for a word and replaces it with main word. This is done with the help of WordNet.

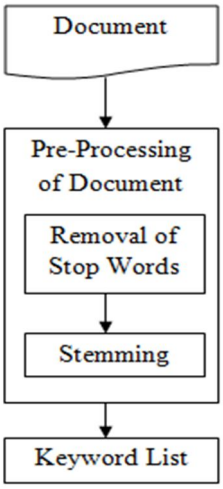

Figure 3. Steps involved in pre-processing

4.1.1 Tf-Idf Weighting

Before applying clustering algorithm on a set of news articles as word documents, for comparing the documents, they must be converted into vector representation. The pre-processed document must be represented with TF-IDF score. TF-IDF stands for Term Frequency-Inverse Document Frequency which results the importance of a term among the document. Term frequency is calculated by dividing the number of occurrences of a word in its document by total number of word in the document. It is a normalized frequency. Inverse document frequency is calculated by taking log of number of documents to be clustered divided by number of documents containing the term. It gives higher weight to rare items. Multiplying the two metrics together give TF-TDF weighting which gives importance to terms frequent in the particular document weighted for clustering and rare among the documents that are clustered.

Tf –Idf (term, document)
=Tf (term, document)*Idf (term)

Where Tf is term frequency
Idf is inverse document frequency

4.1.2 Cosine Similarity

As a result of TD-IDF weighting, we have represented each news article in the form of word document as vector models. Next step is to find the similarity between the documents. In our method cosine similarity is used to obtain the distance (similarity) between two documents. It is computed by dividing the dot product of two vectors by the product of their magnitudes. This defines equidimensionality and element- wise comparability of document vectors in vector space. The cosine angle is a good indicator of similarity between the two vectors of the documents.

$$\text{Cosine similarity}(\text{vec\_A}, \text{vec\_B}) = \frac{\text{Dot Product}(\text{vec\_A}, \text{vec\_B})}{|\text{vec A}| * |\text{vec B}|}$$

Where vec_A is vector model of document A
vec_B is vector model of document B

4.3 Clustering

The news article documents are to be clustered to improve the results of information retrieval system in terms of precision or recall. This provides better filtered and adequate result to the user. Clustering methods are made into generic categories: hierarchical agglomerative and partitional clustering. Hierarchical clustering is of two types. One forms a sequence of partition in data that leads n clusters from single cluster (divisive) and another merge clusters based on similarity between clusters (agglomerative). The divisive algorithm starts up with each data point as a cluster. Then it merges the tree node that

shares certain degree of similarity. Thus it needs either cluster similarity or distance measure to split or merge data of different cluster. Agglomerative algorithm involves pair wise joining of clusters. Hierarchical clustering algorithms face difficulties in handling different sized cluster and not suitable for large sized data. Thus we prefer partition algorithm which suits large set of data.

Partitional algorithm defines the number of clusters initially, let k and evaluate the data at once such that sum of distance over their cluster center is minimal. Unlike hierarchical clustering, partitional clustering involves single level division of data. There are various types of partitional clustering algorithms: k-means, k-median and k-medoids. These algorithms differ by the approach of defining cluster centers and not how they represent the clusters k-means algorithm defines its center as mean data vector averaged over all data nodes in the cluster. In k-median the median is calculated for each dimension in data vector. In k-medoids the cluster center is defined as an item with smallest sum of distances to other items in the cluster.

4.1.3 K-Means Clustering

K-means algorithm is an unsupervised learning algorithm which is much efficient than other partition algorithm with better initial centroids .It aims to partition n documents into k clusters in which each document belongs to cluster with nearest mean that is, it groups similar document where each group is known as a cluster. Document in each group establish maximum similarity within its group and maximum diversity with other groups.

Step 1: Initialize parameter k, number of cluster centroids based on number of cluster needed.

Step 2: Data points are assigned to the closest cluster based on the cosine similarity.

Step 3: The position of the centroids are recomputed after assigning all data points are assigned to the cluster.

Step 4: Step 2 and 3 are repeated until cluster converge.

Initially the user has to specify the value of k, desired number of cluster centers. Each data point is assigned to the nearest centroid. Set of points assigned to each centroid is known as cluster. When data points are added the centroid for the cluster is updated based on the added data points

4.4 Keyword Extraction And News Article Retrieval

After the clusters are formed by the clustering algorithm, keywords for each cluster have to be defined for each cluster. In order to define key words list for each cluster, we first select the frequent terms in the cluster by setting threshold. The resultant list is fed to the WordNet, an electronic lexical database that describe each English word as noun, adverb, adjective and verb. It also describe the semantic relationship between the word that is, it is whether its synonym or hyponym. WordNet collect the noun candidates from the keyword list of the cluster and consolidate the set of synonym and hypernym words. Thus keywords and the related synonyms and hyponyms are defined for each cluster. Thus the user queries the cluster database with user defined key phrase. The words in the key phrase are compared with the keyword list of cluster. The cluster with which the key phrase matches is said to contain the required news article.

5 CONCLUSION

Re-enactment of newspaper article proposes an approach to segment news article from newspaper and convert those article into word files. These word files are pre-processed to remove stop words and stemming. This pre-processed word file is converted into vector form by means of TF-IDF weighting. Each document is represented by means of a vector. The similarity between the documents is found out by means of cosine similarity. The documents with more similarity are clustered by means of K-means algorithm. For each cluster formed by k-means algorithm keyword list are generated for making retrieval of article based on user queries efficient.

6 REFERENCES

[1] Wei-Yuan et al, Adaptive Page Segmentation for Color Technical Journals' Cover Image, Image and Vision Computing, 16(1998) 855-877, Elsevier Publication.

[2] Fu Chang et al, Chinese Document Layout Analysis Using an Adaptive Regrouping Strategy, Pattern Recognition 38(2005) 261-271,Pergamon Publication.

[3] Osama Abu Abbas et al, Comparisons between Data Clustering Algorithms, volume 5, No.3, July 2008, The International Arab Journal of Information Technology.

[4] FarzadFarahmandnia et al, A Novel Approach for Keyword Extraction in Learning Object Using Text Mining and WordNet, Volume 03, Issue 1(2013) 01-06,Global Journal of Information Technology.

[5] LiangcaiGao et.al, Newspaper Article Reconstruction Using Ant Colony Optimization and Bipartite Graph, Applied Soft Computing 13(2013) 3033-3046, Elsevier publication.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 3, 169 - 175, 2014, ISSN: 2319–8656

Cloud Computing: Technical, Non-Technical and Security Issues

Wada Abdullahi
Federal College of Education (Technical)
Potiskum
Yobe State, Nigeria

Alhaji Idi Babate
Federal College of Education (Technical)
Potiskum
Yobe State, Nigeria

Ali Garba Jakwa
Federal College of Education (Technical)
Potiskum
Yobe State, Nigeria

Abstract: Cloud Computing has been growing over the last few years as a result of cheaper access to high speed Internet connection and many applications that comes with it. Its infrastructure allows it to provide services and applications from anywhere in the world. However, there are numerous technical, non-technical and security issues that come with cloud computing. As cloud computing becomes more adopted in the mainstream, these issues could increase and potentially hinder the growth of cloud computing. This paper investigates what cloud computing is, the technical issues associated with this new and evolving computing technology and also the practical applications of it. It also addresses and highlight the main security issues in cloud computing.

Keywords: computing; technical; security; cloud; issues

INTRODUCTION

Cloud Computing is a broad term used to describe the provision and delivery of services that are hosted over the internet [24]. The "cloud" in cloud computing comes the diagram representation or symbol that is usually used to depict the internet in network diagrams or flowcharts [24]; [19].

In this type of computing, servers on the internet store information permanently and temporarily cached on the client-side devices – including laptops, desktops, hand-held devices, monitors, sensors, tablet computers etc [19]. The infrastructure allows services to be provided and accessed from anywhere in the world as services are offered through data centres – virtually [19]. Here, the cloud becomes the single access point for customers/users to access services. This means that at any given time, a user has as little or as much of a service provided by the particular service provider. However, the user only needs a computer or appropriate device and Internet connection to gain access to the service(s).

In the cloud computing model, there are three main entities as illustrated in Figure 1 – End Users, Application Providers and Cloud Providers. According to [9], the assemble and organisation of users and computing resources "provides significant benefits to all three entities because of the increased system efficiency and availability."

There are numerous reason why cloud computing has gained interest over the last few years. One reason has been the significant and improved innovations in distributed and virtualisation computing. Additionally, the cost-benefits and access to high-speed internet have also contributed to this accelerated interest.

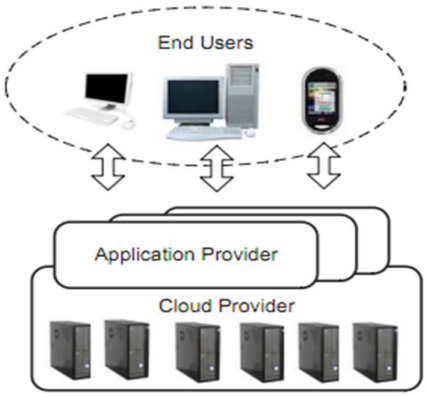

Figure 1: User of a cloud-based application system.

Cloud Providers provide the necessary cloud infrastructure – which includes network facilities, replications sites and data centres. Applications services used on the cloud infrastructure are provided by Application Providers. These application services are then used by the "End Users".

CLOUD COMPUTING MODELS

The model of cloud computing can be divided into private or public. In public cloud, the providers of the cloud sell services to anyone, whereas in private cloud data centres or cloud providers they hosted their services only to a limited or small number of subscribers or buyers. However, there are situations where a service provider uses the resources of a public cloud to provide or create a private cloud – this is what is known as virtual private cloud [24]. In the service provision part of cloud computing, there are three main categories of service provision namely Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS).

In SaaS cloud model, the hardware infrastructure and the provision of the software product is supplied by the service provider supplies. Additionally, the user interacts with the service provider through a front-end portal [24]. In other words, SaaS can be said to be *software that is "owned, delivered, and managed remotely by one or more providers."* [8]. There are many companies that use business applications that are provided remotely [3]. According to Biddick, companies benefit from using SaaS as that development and maintenance of software applications becomes the burden of the provider.

In cloud computing, PaaS is described as "a set of software and product development tools hosted on the provider's infrastructure." This allows developers to develop and create software tools over the internet and also on the provider's platform. The use of APIs (Application Programming Interfaces), a gateway software and website portal on the end user's computer is commonly used by PaaS providers – this allows end users to develop software applications. However, data portability and standards for interoperability are not currently being used in cloud computing [24]. This makes it difficult to systems to work together or easy exchange of information.

In IaaS, providers "provides virtual server instance API to start, stop, access and configure their virtual servers and storage" [24]. End users can pay for the only capacity they require – making it possible to keep cost to a minimum for new start-up businesses. An example of an IaaS is Amazon Web Services – currently the largest provider [15].

SOFTWARE EVOLUTION OF CLOUD COMPUTING

The process of software evolution can be seen as a never-ending process. Once software is developed, it is maintained, and then repeatedly updated with respect to changes in requirements, processes and methodologies. It is known that 90% of companies' software budget is spent on maintenance and adapting existing software tools than developing new ones from scratch [5].

Cloud computing is a specialised distributed computing on a large-scale [7]. However, there are differences from the traditional distrusted systems in many ways;

- ✓ Scalability is massive on cloud computing.
- ✓ Different types or levels of services can be provided to clients outside the cloud with a greater degree of encapsulation.
- ✓ Economies of scale is one of the main drivers of cloud computing.
- ✓ Configuration of services is dynamic and delivery of services can be on demand.

The idea that evolution is driven by change can be observed in cloud computing. There is a growing demand for computing and storage problems in the so called "Internet Age". As a result many companies and individuals are looking to cloud computing to provide the answer [7].

The evolution of cloud computing could be traced back to the 1990s when Grid Computing was used to describe the collection of technologies that enabled users to obtained computing power when required. This lead to the standardisation of protocols to be allows for data exchange over the *grid*. However, according to the commercial utility of grid computing was very limited until about 2007/8 [7]. The vision of both cloud and grid computing technologies remains the same i.e. reduce computing cost with increased reliability, and transform the old style of standalone software computing to one where services can be obtained from third parties via the Internet. The underlying technologies and procedures of cloud and grid computing are somehow different.

Utility computing is a model based on the concept of demand and outsourcing availability [2]. In this type of model, resources and services are provided to the end user and charged based on usage.

The increasing demand for computing comes from our need to analyse large collection of data – data that was not present as of ten years ago. Additionally, there has been the realisation that operating mainframe computers are very expensive compared to commodity clusters. This has lead to a reduced cost of virtualisation. Over the last ten years, companies like Google, Microsoft and Amazon have spent billions of dollars building large-scale computing systems containing a collection of hundreds of thousand computers. The commercialisation of these systems means that computing can be delivered on-demand. The scale of operation of cloud computing is comparatively bigger than that of grid computing. Furthermore, this allows computing to be provided cheaply (economies of scale) than previously thought with grid computing [7].

Cloud computing has evolved through a series of phases – there was the initial grid (or utility) computing phase, then there was the "application service provision" which was then followed what is now known as SaaS [18]. According to Mohammed, the most recent evolution of cloud computing is its development with Web 2.0. This was made possible as bandwidth increased in the late nineties. In 1999, salesforce.com *pioneered the concept of delivering enterprise applications via a simple website*. As a result, companies; both mainstreams and specialists started the delivery of Internet-based services and applications [18]. Following on from that, Amazon Web Services was developed in 2002 – which allowed for many cloud-based services such as computing and storage to be delivered online. In 2006, Elastic Computer Cloud (EC2) was launched by Amazon. EC2 was a commercial web service which enables individuals and small companies to rent computers online to run their own applications [18]. Since 2007, cloud computing has become a "hot topic" due to its flexibility to offer dynamic IT infrastructure and configurable software services over the Internet [25]. The emergence of cloud computing coincides with the development of Virtualisation technologies. Since 2007, Virtualisation technologies have increased and as a result, cloud computing has been observed to have out-paced that of grid computing [26]. This trend still continues to grow as companies and research community propose and develop this computing paradigm – cloud computing. According to Mohammed, a great milestone of cloud computing came about in 2009 with the introduction of Web 2.0. The Web 2.0 is designed to allow the web work as a platform i.e. clients' services do not depend on the operating system (OS) being used by the user [23] The main properties of Web 2.0 are information sharing, user-centred design and interoperability – all of these are factors that have contributed to the continual development of cloud computing over the last few years.

TECHNICAL ISSUES OF CLOUD COMPUTING

The idea behind cloud computing is that software and hardware services are stored in "clouds", web servers rather than a connection of standalone computers over the Internet. Here, a user can access the right services and data they require [2]. Another benefit of cloud computing is that of "moving" data to the cloud to allow for access to a user's data anywhere [2]. An important feature that comes with cloud storage of data is essentially the automation of different management tasks.

It can be noted that a fusion or combination of technologies such as grid computing, autonomic computing (AC) and utility computing has contributed to the evolution of cloud computing. AC is built on the following concepts; self-protection, self-management, healing and configuration. It uses a closed control loop system which allows it monitor and control itself with external input. As the current situation and needs of a system changes, an AC system adapts itself to those dynamical changes – making it self-adaptive as well. This combined with grid computing which was known to be "heterogeneous and geographically detached" [2], has produced a new computer architecture for cloud computing.

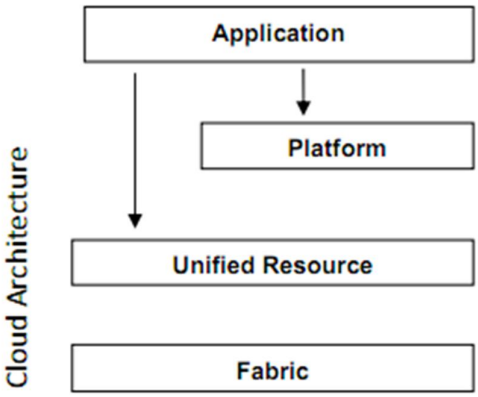

Figure 2: The Cloud Computing Architecture

[7]define a four-layer architect of cloud computing (Figure 2). These layers are Application, unified resource, platform and fabric. The physical hardware resources (such as storage resources, computing resources and network resources are contained in the *Fabric layer*. Abstracted/encapsulated resources (usually as a result of virtualisation) are contained in the *Unified Resource layer*. These abstracted resources are exposed to the upper layer and also by the end user as integrated resources such as a database system, logical file system, a virtual cluster/computer [7]. Specialised tools and technologies such as middleware and other services are provided by the *Platform layer* in addition to the resources already contained in the unified resource layer to provide a platform for the development and deployment of applications. Lastly, applications that will run in the cloud are contained in the *Application layer*. There are three different levels of services that are provided by cloud computing (IaaS, PaaS and SaaS). The type of service provision depends on the layer which the service provider wants to make available. However, it is also possible for a service provider to expose services on more than one layer.

IaaS allows for the provision of software, hardware and equipment usually at the unified resource layer. PaaS provides users with a *high-level integrated environment to build, test, and deploy* their own built applications. Here, developers are faced with certain technical limitations and restrictions on the type of software tool or application that they can develop in *exchange for built-in application scalability* [7]. The Google App Engine now enables users to build web application by using the same systems that Google uses to run its application. SaaS, on the other hand, enable service providers to provide specifically built applications and software that can be accessed remotely by end users via the Internet using a utility or usage-based model of pricing [7].

Here, the main security issue is that of the openness of the hosting or services providers. Building a test environment in the cloud requires hosted compilers which can provide a gateway for "hacker" and experience programmer alike to develop and deploy malicious programs in the cloud. A possible way to deal with this security threat for cloud provides to accept pre-compiled programs that are thoroughly scanned for viruses before being deployed. This can be achieved by restricting users to deploying programs only in the application layer – thereby restricting the risk of contamination across layers within the cloud.

Furthermore, the provision of services at different levels brings about the need for standards to be defined to allow for the exchange of information and services between clouds. To date, such standards do not exist that cause interoperability issues. There is always a security concern when standards and protocols are not properly defined in any computing environment. As cloud continues to grow and mature, there will be the need to adopt industry-wide standards that will ease interoperability issues and increase the levels of services that cloud provider can deliver to end users.

There are numerous security concerns when it comes to software development over cloud computing. This makes it difficult for certain computing techniques to be incorporated in software development. Additionally, some techniques make programs or software vulnerable in distributed systems and in this scenario, cloud computing.

CROSS-CUTTING CONCERN

Cross-cutting concerns in software development relates to aspects of a program that affect or crosscut other modules or concerns [1]. Usually, these concerns arise due to difficulty in decomposing them from a program in the developmental stage, which includes the design, coding and implementation phases, as a result can occur in the duplication of code (known as scattering) or tangling (these come about when systems have significant dependence on each other) or both . Some examples of cross-cutting concerns include:

- Exception handling
- Validation
- Logging
- Authentication and authorisation

A suggested way to deal with cross-cutting concerns in program development is to use Aspect-Oriented Programming (AOP). Aspects relates to a feature or part of a program that is not linked to the core functionality of the program but linked to many other parts of the program. Using separation of concerns (SoC), cross-cutting can be

reduced or eliminated from program development – this is the basis of AOP. SoC is a process of distinctly separating functions of a program to avoid or limit functionality overlapping. Traditionally, SoC was achieved mainly by encapsulation and modularisation. AOP aims to increase modularity by enabling SoC. It also requires programs to be broken in distinct logical parts – SoC called concerns.

Developing programs on cloud computing can be done using AspectJ – a java extension which is known as the de facto standard development tool for AOP- to ease cross-cutting worries [16]. It can be challenging to develop a program on cloud as it might difficult to ascertain how to break down a program into logical parts on different servers.

PROGRAM SLICING

Another technique of software development is that of program slicing. Program slicing relates to the simplification of programs *by focusing on selected aspects of semantics*. The process of program slicing involves the deletion of certain parts of a program that are known to have no impact or effect on a particular semantics. Program slicing also allows developers to focus more attention of the parts of the program that can cause a fault. As a result, there are application of program slicing in testing and debugging, program comprehension, software re-engineering and measurement [10].

There are mainly two dimensions to program slicing; the semantic dimension and the syntactic dimension. The preservation of parts of the program relates to the semantic dimension. Here, the static behaviour of the program is unaffected after slicing and likewise, dynamic criteria enable the dynamic behaviour of the system to be preserved. Under the semantic dimension, slicing can be dynamic, static or conditioned [10]. However, there is less choice under the syntactic dimension. Here, there are two main possibilities; firstly the syntax of an original program is preserved, where possible, by moving parts of the programs which does not affect the interested semantic, secondly program slicing is freely allowed to perform any syntactic transformation that preserves semantic conditions – this is known as amorphous slicing [10].

Program slicing could have issues with regards to cloud computing. Deleting of certain parts of the program on clouds can affect other applications. Additionally, parts of a program that are thought of having no impact on core semantic of a program on one server could have a bigger impact on a program on another server.

PROGRAM OR APPLICATION CLUSTERING

Clustering, in computing, relates to a group of computers or servers dedicated to performing a single task. Software systems are used to configure servers to cluster in application clustering. Servers are connected together by a software program which enables the servers to perform individual tasks like failure detection and load balancing [4]. Here, applications are installed individually on the servers and are pooled in together to handle various tasks when required. It becomes important for the cluster to effectively handle routing of data to and from the cluster [4].

In cloud computing, program clustering helps achieve scalability – the ability of the cloud to appropriate resources to specific tasks i.e. when a task needs more computing resources, it has the ability to recruit more servers or computing power to perform that specific task. The benefit of cloud computing is that it contains hundreds of thousands of connecting computers which makes it easy to distribute work load. There are symmetric clusters where workload is distributed evenly amongst the clustering servers and asymmetric clusters have the ability to reserve particular servers only for use when the main servers fail. Cloud computing provides a single point of access to its end users to gain access to application services stored on the servers in the "cloud". Servers can fail; as a result clouds must tackle the issue of passing tasks around when servers failed.

NON-TECHNICAL ISSUES OF CLOUD COMPUTING

Cloud computing comes with other non-technical issues or concerns which if not tackled could restrict the growth and evolution of cloud computing.

INADEQUATE SECURITY

Most cloud vendors support what is known as multi-tenancy compute environment by design. What is most important is that, vendors must decide on the right balance between providing essential infrastructure and internal security and the quest for improved cloud computing services. According to [27], trustworthiness is important when it comes to SaaS services. With SaaS, data privacy and security are the most important factors for end users (also known as tenants).

LACK OF COMPUTABILITY WITH EXISTING APPLICATIONS

Another major issue currently facing cloud computing is the lack of inherent computability with existing applications. There are, however, efforts to change this. What is observed in order to improve scalability and improve the level of services provided to users, vendors are now providing snippets of existing codes in the case of PaaS. What this means is that new applications are becoming cloud-specific.

LACK OF INTEROPERABILITY BETWEEN CLOUDS

The lack of a standardisation across platform increases cost of switching clouds and also increases the complexity of code in the event of program migration. Since cloud vendors have different application models, there are vertical integration problems which make it virtually impossible at time to move from one cloud to another. As this is major issue, a user has to be careful when choosing the right vendor to obtain services from.

OTHER ISSUES

There is also the issue of service legal arrangement which prohibits a user from moving from one cloud to another unless certain conditions are met. This increases switching costs for the end user and subsequently, gives more power to the cloud vendor.

LEGAL ISSUES

According to [17], the biggest issue concerning cloud computing comes from governments. This is a result of the borderless global network operations of cloud computing. Unlike grid computing, cloud computing is not geographic-specific. Having no borders makes it difficult for

governments to protect or control how data of its people is stored or used elsewhere and also how to tax companies operating services over a cloud. Under taxation, if a company is taxed based on geographical location of its computing operation, it can simply move this to a virtual office in a country with a lower tax rate [17].

There are measures being taken to tackle the issue of taxation under cloud computing on a global approach in order to stop companies from exploiting tax advantages. Additionally, there is a recognised need for *harmonised laws* in the global front to police how data is stored and used over the cloud.

SECURITY ISSUES

According to [11], one of the main security concerns of cloud computing is that of its immaturity of the technology. The cloud provider and client have both security responsibilities depending on the type of service. In the case of an IaaS service model, the virtualization software security, environment security and physical security rest with the cloud provider. The client or user is responsible for operating system, data and applications. However, in a SaaS model, software services, physical and environment security are the responsibility of the cloud provider.

The main security concern with cloud computing is that of data security. Confidential documents stored on the cloud can become vulnerable to attacks or unauthorised access or modification. There is also the issue of where the data is physically stored i.e. where the data stores are located. Some companies prohibit the storage of their data in certain jurisdictions or countries [13]. Here, trust in cloud computing is very vital in ensuring that data is properly managed as depending on the type if model adopted, IaaS, SaaS or PaaS, the governance of applications and data lies outside the control of the owner [6]. A possible address to this security issue is to use an Active Directory (AD or LDAP) in authenticating users who can have access to data and applications in the cloud. Using Access Control Lists (ACLs), permissions can be assigned per document stored and restrict users from unauthorized access and modification. Additionally, there are now various security software tools which can be deployed in the application layer to provide and enhance authentication, operational defence, confidentiality and message integrity [12].

There are numerous encryption techniques that have been developed to further to ensure that data is securely stored in the cloud. In [14], the authors used an encryption technique known as “Elliptic curve cryptography encryption” in an attempt to protect and make secure, data stored in the cloud. According to [20], clouds are constantly being attacked, on a daily basis, and as such extra security protocols are needed to ensure security integrity. The authors proposed the use of “Transparent Cloud Protection System (TCPS)” to increase cloud security. Although, their system provided increased virtualization and transparency, it was never tested in a professional cloud environment and as such makes it difficult to establish how useful such a system really is.

Another possible way to address security issues in the cloud is to use Trusted Third Part (TTP) services within a particular cloud. Here, TTP established trusted and secure interaction between trusted parties in the cloud. Any untrusted party, can simply be ignored or blocked from access data and application within that cloud. TTP can be useful in ensuring confidentiality; authenticity and integrity are maintained in the cloud [6].

The major security worry is that most concerns and issues discussed in this review are looked at the problems in isolation. However, according to [22] the technical and security issues of cloud computing need to analysed together to gain a proper understanding of the widespread threat of the “cloud”.

APPLICATIONS OF CLOUD COMPUTING

One of the reasons for the upward trend of resources committed to cloud computing is that, cloud computing has many real benefits and applications to companies, individuals, research bodies and even government. As the size of data increases, the need for computing power capable of analysing these data increases relatively.

One application of cloud computing is that clients can access their data and applications from anywhere at any particular time. Additionally, clients only need a computing device and an Internet connection to access their data and applications. For example, Dropbox [21] allows users to store their data on an online cloud and access it using any computing device. Users can also share folders with other users in the same manner. Another example is Google Docs [21] which allows users to edit, modify documents online without having to move the documents around. All modifications are saved on the master document on the cloud.

Cloud computing has the possibility of reducing hardware costs for many companies and individuals. Clients can gain access to faster computing power and bigger storage without paying for the physical hardware.

Cloud computing gives companies the ability to gain company-wide access to its host of software or applications. Here, a company does not have to buy a licence for every employee to use particular software; rather it can pay a cloud computing company on a usage fee basis (utility computing model). With the introduction of Web 2.0, access to cloud computing has become less OS-dependent. [21].

CONCLUSION

Cloud Computing describes the provision and delivery of services that are hosted over the internet according to [24]. The infrastructure of cloud computing allows services and applications to be provided and accessed from anywhere in the world as services are offered through data centres [19]. Here, the cloud becomes the single access point for customers/users to access services.

The number of reasons have contributed to the success of cloud computing over the last few years. One reason being the significant improvement of innovations in distributed and virtualisation computing. Furthermore, cheaper access to high-speed internet has also contributed to this accelerated interest [18].

There are three main categories of service provision in cloud computing are Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS). In SaaS cloud model, the hardware infrastructure and the provision of the software product is supplied by the service provider supplies. PaaS is described as “a set of software and

product development tools hosted on the provider's infrastructure." Developers are able to develop and create software tools over the internet and also on the provider's platform using APIs providers by the cloud provider. IaaS providers "provides virtual server instance API to start, stop, access and configure their virtual servers and storage" [24].

However, there are technical issues that need to be considered carefully when cloud computing is concerned – one of them being the use of program clustering. Program clustering helps achieve scalability – the ability of the cloud to appropriate resources to specific tasks i.e. when a task needs more computing resources, it has the ability to recruit more servers or computing power to perform that specific task. This requires the need for a management system in recruiting servers and one that is able to prevent bottlenecks when servers fail. Program slicing allows for the simplification of programs *by focusing on selected aspects of semantics*. Deleting parts of a program that are thought of having no impact on core semantic of a program on one server could have a bigger impact on a program on another server. It can be challenging to develop a program on cloud as it might difficult to ascertain how to break down a program into logical parts on different servers.

In addition to technical issues, there are real security issues with regards to cloud computing. The most concerning security issue is that of data privacy and integrity in the cloud. However, there are many works and techniques being developed to combat the threat of data and application misuse and access in the cloud. The issues, both technical and security related, have all been observed in isolation. What must be noted is that, when these issues are combined, it could be a huge threat to cloud computing and such it is imperative that the issues be addressed not in isolation.

One of the reasons for the upward trend of resources committed to cloud computing is that, cloud computing has many real benefits and applications to companies, individuals, research bodies and even government. As the size of data increases, the need for computing power capable of analysing these data increases relatively. Another application of cloud computing is that user can access their data and applications from anywhere at any particular time.

As cloud computing continues to grow, so we hope the technical issues and as the need for standardisation to allow clouds to exchange information effectively and concisely. Also the current well established cloud vendors to be prepared to get rid of their standards.

ACKNOWLEDGEMENTS

Our thanks to the colleague Lecturers of Computer Science department Federal College of Education (Technical) Potiskum for their contributions towards development of the paper.

REFERENCES

[1] Abdullin, R. (2010) "Cross-cutting concern". From: http://abdullin.com/wiki/cross-cutting-concern.html, Accessed 11th Feb 2013.

[2] Aymerich, F., Fenu, G. and Surcis, S. (2008). "An Approach to a Cloud Computing Network". First International Conference on the Applications of Digital Information and Web Technologies, ICADIWT.

[3] Biddick, M. (2010). "Why You Need a Saas Strategy". Retrieved February 12, 2013, from: http://www.informationweek.com/news/services/saas/showArticle.jhtml?articleID=222301002

[4] Bliss, H. (2010). "What is Application Clustering?" Available at: http://www.wisegeek.com/what-is-application-clustering.htm, Accessed 14th Feb 2013

[5] Brooks, F. (1997). "The Mythical Man-Month". Addison-Wesley.

[6] Dimitirios, Z. and Dimitrios, L. (2012) Addressing cloud computing security issues. *Elsevier*, 28 (3), p.583–59, Available at: http://www.sciencedirect.com.ergo.glam.ac.uk/science/article/pii/S0167739X10002554. Accessed: 10thFeb, 2013.

[7] Foster, I., Zhao, Y., Raicu, I. and Lu, S. (2008) "Cloud Computing and Grid Computing 360-Degree Compared" [Online Article] from: http://arxiv.org/ftp/arxiv/papers/0901/0901.0131.pdf, Retrieved 14 Feb 2013.

[8] Gaw, P. (2008). "What's the Difference between Cloud Computing and SaaS". Available at: http://cloudcomputing.sys-con.com/node/612033, Accessed on 11 Feb 2013.

[9] Gu, L. and Cheung, S-C. (2009). "Constructing and Testing Privacy-Aware Services in a Cloud Computing Environment – Challenges and Opportunities". *Internetware*. ACM 978-1-60558-872-8/10.

[10] Harman, M. and Hierons, R.M. (2006). "An Overview of Program Slicing". Available at: http://www.cs.ucl.ac.uk/staff/mharman/sf.html, Retrieved on: 10th Feb. 2013.

[11] Hocenski, Ž. and Kresimir, P. (2010) "Cloud computing security issues and challenges ", paper presented at *MIPRO, 2010 Proceedings of the 33rd International Convention*, 24-28 May. IEEE Conference Publications, p.344 – 349. Available at: http://ieeexplore.ieee.org.ergo.glam.ac.uk/stamp/stamp.jsp?tp=&arnumber=5533317, Accessed 15th Feb. 2013.

[12] Karadesh, L. (2012) Applying Security Policies and service level Agreement to Iaas service Model to Enhance Security and Transition. *Elsevier*, 31 (3), p.315–326. Available at: http://www.sciencedirect.com.ergo.glam.ac.uk/science/article/pii/S0167404812000077, Accessed: 12th Feb, 2013.

[13] King, N. and Raja, V. (2012) Protecting the privacy and security of sensitive customer data in the cloud. *Elsevier*, 28 (3), p.308–319. Available at: http://www.sciencedirect.com.ergo.glam.ac.uk/science/article/pii/S0267364912000556 , Accessed: 10th Feb, 2013.

[14] Kumar, A. et al. (2012) "Secure Storage and Access of Data in Cloud Computing", paper presented at *ICT Convergence (ICTC), 2012 International Conference on*, 15-17th Oct.. IEEE Conference Publications, p.336 - 339.

Available at: http://ieeexplore.ieee.org.ergo.glam.ac.uk/stamp/stamp.jsp?tp=&arnumber=6386854, Accessed: 12th Feb, 2013.

[15] Lewis, C. (2009). "Infrastructure as a Service". Available at: http://clouddb.info/2009/02/23/defining-cloud-computing-part-6-iaas/, Accessed: 6th Feb 2013.

[16] Li, S. (2005). "An Introduction to AOP". Available at: http://www.ibm.com/developerworks/java/tutorials/j-aopintro/section4.html [Accessed 12th Feb 2013].

[17] Lonbottom, C. (2008) "Obstacles to Cloud Computing". [Online] Available at: http://www.information-management.com/news/10002177-1.html?pg=1, Accessed 11th Feb 2013.

[18] Mohammed, A. (2009). "A history of cloud computing", Available at: http://www.computerweekly.com/Articles/2009/06/10/235429/A-history-of-cloud-computing.htm, Retrieved: 7th Feb 2013.

[19] Schneider, L. (2011). "What is cloud computing?" Available at: http://jobsearchtech.about.com/od/historyoftechindustry/a/cloud_computing.htm Accessed on 10th Feb 2013.

[20] Shaikh, F. and Haider, S. (2011) "Security threats in cloud computing ", paper presented at *6th International Conference On Internet Technology And Secured Transactions*, Abu Dhabi, 11-14th Dec. IEEE Conference Publications.

[21] Strickland, J. (2011) "How Cloud Computing Works" [Online] Available http://computer.howstuffworks.com/cloud-computing2.htm Accessed 14th Feb 2013.

[22] Sun, D. et al. (2011) Addressing cloud computing security issues. *Elsevier*, 15 p. 2852–2856. Available at: http://www.sciencedirect.com.ergo.glam.ac.uk/science/article/pii/S1877705811020388 Accessed: 12th Feb, 2013.

[23] TechPluto (2009). "Core Characteristics of Web 2.0 Services". Available http://www.techpluto.com/web-20-services/ [Accessed 10th Feb 2013].

[24] TechTarget (2007). "Cloud Computing" Retrieved from: http://searchcloudcomputing.techtarget.com/definition/cloud-computing, Accessed on: 12th Feb 2013.

[25] Wang, C. et al. (2008) "Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing", paper presented at *INFOCOM, 2010 Proceedings IEEE* , IEEE Conference Publications, p.1-9.[online] Available at: http://ieeexplore.ieee.org.ergo.glam.ac.uk/stamp/stamp.jsp?tp=&arnumber=5462173, Accessed on: 12th Feb 2013.

[26] Wang, L. et al. (2018) Cloud Computing: A Perspective Study. *New Generation Computing, Springer Link*, 28 (2), P.137-146.

[27] Zhang Q., Cheng L., and Boutaba R., (2009) Cloud Computing: State of the art and research challenges J internet Serv Appl 1 Brazilian Computer Society pg. 7 – 18.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 3, 176 - 179, 2014, ISSN: 2319–8656

Stepping Stone Technique for Monitoring Traffic Using Flow Watermarking

S.R. Ramya
Department of CSE
PPG Institute of Technology
Coimbatore, Tamilnadu, India

A. Reyana
Department of CSE
PPG Institute of Technology
Coimbatore, Tamilnadu, India

Abstract : The proposed system describes a watermarking technique on ownership authentication providing secured transactions. The unique watermark signature is invisible. The specific request preferred by the user is identified by the watermark extraction procedure, which identifies the signature and returns the user requested data with a proper secret key, indicating authorized user. The watermark extraction algorithm returns an error that tells impostor user. Here it requires a unique signature during both the insertion and the request procedures, thus the user remains unauthorized until it passes the signature validation test. Here the versions of signature and secret key techniques are followed.

Keywords – Perturbation, Embedding, Correlation, Extraction, validation

1. INTRODUCTION

Today, creators and owners of digital video ,audio, document and images fears to put their multimedia data over the Internet, because there is no way to track the illegal distribution and violation of protection. Without mechanisms to support the above requirements, owners cannot generate proof that somebody else violated law. The techniques that have been proposed for solving this problem are collectively called unique digital watermarking. Unique digital watermarking refers to the embedding of unobtrusive marks or labels that can be represented as bits in digital content. The method also provides a unique way for propagating information in the form of an encrypted document. Existing connection correlation approaches are based on three different characteristics: 1) host activity; 2) connection content; and 3) inter-packet timing characteristics. The host activity based approach collects and tracks users login activity at each stepping stone, therefore not trustworthy as the attacker is assumed to have full control over each stepping stone, he/she can easily modify, delete or forget user login information. Content based correlation approaches require that the payload of packets remains invariant across stepping stones. And the attacker can easily transform the connection content by encryption at the application layer; these approaches are suitable only for unencrypted connections. The traffic timing based approaches monitors the arrival or departure times of packets, and uses this information to correlate incoming and outgoing flows of a stepping stone.

2. PROPOSED SYSTEM

The proposed system has a robust technique that is unique watermarking and image authentication schemes. The proposed scheme includes two parts. The first is a unique watermarking which will be embedded into image for ownership authentication. The second is a signature verification process, which can be used to prove the integrity of the image. The unique signature will be extracted from the image. The signature is verified when the image is incidentally damaged such as loss compression thus provides a high degree of robustness against the attacker, the attacker can add the secret key in watermarking, which can be easily analyzed to identify the intruder. Thus all the packets in the original flow are kept and no packets are dropped from or added to the flow by the stepping stone. Attackers commonly relay their traffic through a number of (usually compromised) hosts in order to hide their identity. Detecting such hosts, called stepping stones, is therefore an important problem in computer security. The detection proceeds by finding correlated flows entering and leaving the network. Traditional approaches have used patterns inherent in traffic flows, such as packet timings, sizes, and counts, to link an incoming flow to an outgoing one rather than storing or communicating traffic patterns, all the necessary information is embedded in the flow itself. This, however, comes at a cost: to ensure robustness.

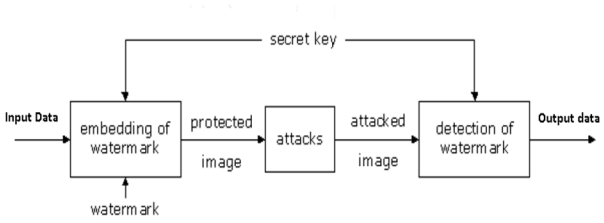

Fig 1. Correlation Analysis

3. SYSTEM DESCRIPTION

3.1 Watermark Bit Embedding And Decoding

Watermarking bit embedding involves the selection of a watermark carrier embeds with unique watermark signature. At the time of user registration, it collects the unique watermarking signature from the user. This process embeds the signature by a slight modification of some property of the carrier. The embedded bit watermark is guaranteed to be not corrupted by the timing perturbation. The watermark is subsequently embedded by delaying the packets by an amount such that the IPD of the watermarked packet.

The IPD is conceptually a continuous value; it first quantizes the IPD before embedding the watermark bit. Given any IPD ipd > 0, we define the quantization of ipd with uniform quantization step size s > 0 as the function q (ipd, s) = round (ipd/s) - - (1) where round(x) is the function that rounds off real number x to its nearest integer. The quantization for scalar x. It is easy to see that q (k s, s) = q (k s + y, s) for any integer k and any y [-s/2, s/2). Let ipd denote the original IPD before watermark bit w is embedded, and ipdw denote the IPD after watermark bit w is embedded. To embed a binary digit or bit w into an IPD, we slightly adjust that IPD such that the quantization of the adjusted IPD will have w as the remainder when the modulus 2 is taken. Given any ipd > 0; s > 0 and binary digit w, the watermark bit embedding is defined as function e (ipd; w; s) = [q(ipd + s=2; s) + ¢] £ s (2) where ¢ = (w ¡ (q(ipd + s=2; s) mod 2) + 2) mod 2. The embedding of one watermark bit w into scalar ipd is done through increasing the quantization of ipd + s=2 by the normalized difference between w and modulo 2 of the quantization of ipd+s=2, so that the quantization of resulting ipdw will have w as the remainder when modulus 2 is taken. The reason to quantize ipd+s=2 rather than ipd here is to make sure that the resulting e(ipd;w; s) is no less than ipd, i.e., packets can be delayed, but cannot be output earlier than they arrive. The embedding of watermark bit w by mapping ranges of unwatermarked ipd to the corresponding watermark ipdw. The watermark bit decoding function is defined as d (ipdw; s) = q (ipdw; s) mod 2.

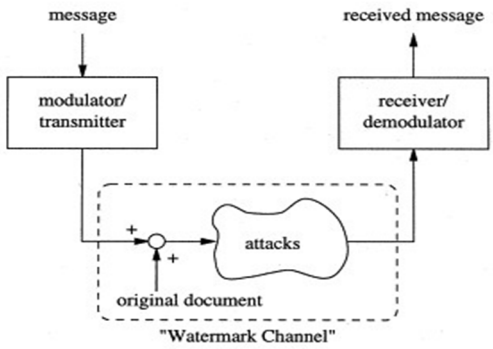

Fig 2. Tracing Model

3.2 Watermark Tracing Model

The watermark tracing approach exploits the observation that interactive connections are bidirectional. The idea is to watermark the backward traffic of the bidirectional attack connections by slightly adjusting the timing of selected packets. If the embedded watermark is both robust and unique, the watermarked back traffic can be effectively correlated and traced across stepping stones, which has not gained full control on the attack target. The attack target will initiate the attack tracing after it has detected the attack. Specifically, the attack target will watermark the backward traffic of the attack connection, and inform sensors across the network about the watermark. The sensors across the network will scan all traffic for the presence of the indicated watermark, and report to the target if any occurrences of the watermark are detected. Gateway, firewall and edge router are good places to deploy sensors, deployed based on the administrative privilege. Since the backward traffic is watermarked at its very source - the attack target, which is not controlled by the attacker. The attacker will not have access to an unwatermarked version of the traffic. This makes it difficult for the attacker to determine which packets have

been delayed by the watermarking process, running at the target.

3.3 Correlation Analysis And Decoding

The number of packets available is the fundamental limiting factor to the achievable effectiveness of our watermark based correlation. This compares and evaluates the correlation effectiveness of our proposed active watermark based correlation and previous passive timing-based correlation under various timing perturbations. By embedding a unique watermark into the inter-packet timing, with sufficient redundancy, we can make the correlation of encrypted flows substantially more robust against random timing perturbations. We can correlate the watermark signatures and identify it's the positive or negative correlation, if positive occurs it detect it is the authenticated user otherwise, if negative occurs it detect it is an Intruder.

To map parameter with Secret Key, we generate secret key and add them into decrypt response. The parameter mapping does not affect the effectiveness of lossless recoverability. Finally the authenticated user takes the requested file in zip format with proper password. Finally the packet header information is extracted for analysis. Packet contents are decrypted in the analysis process. Watermark, source and time information are extracted from the packets. Address verification is also carried out in the packet analysis. The source information is verified in the user authentication process. User information is maintained in encrypted form. Watermarks are used to represent user identity. Time information is also used in the user authentication process.

3.4 WATERMARKING AND EXTRACTION

Flow watermarking is used in the authentication process. Watermarks are embedded by the source node and the receiver node verifies the watermarking images that are updated in the packets. An invisible watermark must be perceptually unnoticeable. Adding the watermark should not corrupt the original audio, video, or image. An invisible watermark should also be robust to common signal distortions and the removal of the watermark should result in degradation of the quality of the original digitized medium. Moreover, the watermark should serve as an original signature of the owner, so that retrieving the watermark from a digitized medium would readily identify the original owner. In order to extract the watermark, both the original image and the watermarked image are needed. First, DCT of the entire watermarked image is computed to obtain the image spectrum. Then, the DCT of the original image is computed. Next, the difference between the two spectrums is computed to extract the watermark X*. Finally, the originally watermark X is compared with the extracted watermark using the following equation: sim (X, X*) = (X X*) / sqrt (X X*). If the original watermark is similar to the extracted watermark, then the watermarked image belongs to the original owner.

Fig 3. Watermarked image

Fig 4. Original image

4. CONCLUSION AND FUTURE SCOPE

The watermarking of multimedia image prevents unauthorized copies from being distributed without the consent of the original owner. Stepping stones are used to hide identity and origin of the attacker. Flow watermarking technique is used to detect attacks with encrypted packets and time perturbed data. The system is enhanced to perform detection with minimum test packet count that manages the detection of stepping stone attacks. Time information is used in the delay analysis. Time information is perturbed in the header. Transmission delay is verified in the system. Packet modification is identified in the delay analysis. The system improves the detection rate.

5. REFERENCES

[1]A. Blum, D. Song, and S. Venkataraman, Detection of Interactive Stepping Stones: Algorithms and Confidence Bounds, *Proceedings of the 7th International Symposium on Recent Advances in Intrusion Detection (RAID 2004). Springer, October 2004*

[2]R. C. Chakinala, A. Kumarasubramanian, R. Manokaran, G. Noubir, C. Pandu Rangan, and R. Sundaram. Steganographic Communication in Ordered Channels, *Proceedings of the 8th Information Hiding International Conference (IH 2006), 2006*

[3]I. Cox, M. Miller, and J. Bloom. Digital Watermarking. *Morgan- Kaufmann Publishers, 2002.*

[4]P. Danzig, S. Jamin, R. Cacerest, D. Mitzel, and E. Estrin. An Empirical Workload Model for Driving Wide-Area TCP/IP Network Simulations. *Journal of Internetworking, 3(1) pages 1–26, March 1992.*

International Journal of Computer Applications Technology and Research
Volume 3– Issue 3, 180 - 185, 2014, ISSN: 2319–8656

A Survey of Web Spam Detection Techniques

Mahdieh Danandeh Oskuie
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: Internet is a global information system. Most of the users use search engines due to high volume of information in virtual world in order to access to required information. They often observe the results of first pages in search engines. If they cannot obtain desired result, then they exchange query statement. Search engines try to place the best results in the first links of results on the basis of user's query.

Web spam is an illegal and unethical method to increase the rank of internet pages by deceiving the algorithms of search engines. It involves commercial, political and economic applications. In this paper, we firstly present some definitions in terms of web spam. Then we explain different kinds of web spam, and we describe some method, used to combat with this difficulty.

Keywords: HITS; Machine learning; PageRank; Search Engine; Web pam.

1. INTRODUCTION

Nowadays, with regard to increasing information in web, search engines are considered as a tool to enter the web. They present a list of results related to user query. A legal way to increase sites rank in the list results of search engines is increasing the quality of sites pages, but this method is time-consuming and costly. Another method is use illegal and unethical methods to increase the rank in search engines. The effort of deceiving search engines is called web spam.

Web spam has been considered as one of the common problems in search engines, and it has been proposed when search engines appeared for the first time. The aim of web spam is to change the page rank in query results. In this way, it is placed in a rank higher than normal conditions, and it is preferably placed among 10 top sites of query results in various queries.

Web spam decreases the quality search results, and in this way it wastes users' time. When the number of these pages increases, the number of pages investigated by crawlers and sorted by indexers increases. In this case, the resources of search engines are lost, and the time of searching in response to user query increases.

According to a definition presented by Gyongyi and Garcia, it refers to an activity performed by individuals to increase the rank of web page illegally[1]. Wu and et al. have introduced web spam as a behavior deceiving search engines [2].

The successes that have been achieved in terms of web spam decrease the quality of search engines, and spam pages are substituted for those pages whose ranks have increased by using legal method. The negative effect of increasing the number of pages spam in internet has been considered as crucial challenge for search engines [3]. It reduces the trust of users and search engine providers. Also, it wastes computing resources of search engines [4]. Therefore, if an effective solution is presented to detect it, then search results will be improved, and users will be satisfied in this way.

Combatting with web spam involves web spamming detection and reducing its rank while ranking or its detection depending on the type of policy [5].

2. VARIOUS KINDS OF WEB SPAM

The word "spam" has been used in recent years to point to unwanted and mass (probably commercials) massages. The most common form of spam is email spam. Practically, communication media provide new opportunities to send undesired messages [6].

Web spam has been simultaneously emerged with commercial search engines. Lycos is the first commercial search engine, and has emerged in 1995. At first, web spam was recognized as spamdexing (a combination of spam and indexing). Then, search engines tried to combat with this difficulty [5]. With regard to article presented by Davison in terms of using machine learning methods to detect web spam, this subject has been taken into account as a university discussion [7]. Since 2005, AIRWeb[1] workshops have considered a place for idea exchanging of researchers interested in web spam [5].

Web spam is the result of using unethical methods to manipulate search results [1, 8, 9]. Perkins has defined web spam as follows: "The attempt to deceive algorithms related to search engines" [9].

Researcher have detected and identified various type of web spam, and they have been divided into three categories:

- Content based spam
- Link based spam
- Page-hiding based spam

[1] Adversarial Information Retrieval on the Web

2.1 Content-based web spam

Content-based web spam has changed the content of page to obtain higher rank. Most of content spamming techniques target ranking algorithms based on TF-IDF. Same of the methods used in this spam is as follows [1]:

- *Body spam:*
  One of the most popular and the simplest methods of spamming is body spam. In this method, terms of spam are placed in documents body.
- *Title spam:*
  Some search engines consider higher weights for the title of documents. Spammers may fill in this tag with unrelated words. Therefore, if higher weight is dedicated to the words of tag from search engine, then the page will receive higher rank.
- *Meta tag spam:*
  The HTML meta tag explanations allow the page designer to provide a short explanation about the page. If unrelated words are placed here, and search engine algorithms consider these pages on the basis of these explanations, then page will receive higher rank for unrelated words. Nowadays, search engines consider lower performance to this tag or ignore it.
- *URL spam:*
  Some search engines break URL of a web page into the terms, sometimes; spams create long URLs containing spam terms. For example, one of URLs created by this method is follows:

  Buy-canon-rebel-20d-lens-case.camerasx.com
- *Anchor text spam:*
  Like document title, search engines dedicate higher weight to anchor text terms, and it presents a summary about the document to which is pointed. Hence, spam terms are sometimes placed in anchor text of a link.
- *Placing spam terms into copied contents:*
  Sometimes, spammers copy the texts on web, and place spam terms in random places.
- *Using many unrelated terms:*
  Spammers can misuse these methods. The page that has been created by this spamming method is displayed in many query words.
- *Repetition of one or more special words:*
  Spammers can obtain high rank for considered page by repeating some the key words. If ranking algorithms of search engines it will be effective.

2.2 Link-based web spam

Link-based web spam is manipulation of link structure to obtain high rank. Some of them have been mentioned as follows[10]:

- *Link farm:*
  Link farm is a collection of pages or sites connected to each other. Therefore, each page will have higher link by creating link farms.
- *Link exchange:*
  Web site owners help each other to add a link to your site. Usually, web site owners obviously show this intention on web pages, or they may be sent to other site owners to request link exchange.
- *Buying the link:*
  Some owners of web sites buy their own web sites from other sites providing this service.
- *Expired domains:*
  Spammers buy expired domains, and unused content is placed over it. Some expired domains may not be already admired, and the links of other sites may remain in these domains, and the validity of those domains is misused.
- *Doorway pages:*
  Web pages involve links. Usually links in this doorway page point to the page of web site. Some spammers may create many doorway pages to obtain higher rank.

2.3 Page-hiding based web spam

Page hiding-based web spam presents a different content to search engines to obtain high rank. Two samples have been mentioned here [11]:

- *Cloaking:*
  Some web sites present different content to search engine rather than to users. Usually, web server can detect and identify company's robots of search engines by IP address, and sends a content different form a page presented to normal users.
- *Redirection:*
  Main page uses different web spamming techniques to be seen by the search engine. When a user refers to a page through search result link, redirection is performed during loading a page.

3. The METHODES OF COMBATTING WITH WEB SPAM

The experts of search engine combat with web spam methods, and they have presented various methods to combat with it, Such as machine learning method and link-based algorithms.
In machine learning method, the classifier predicts that whether the web page or web site has spam or not. This is predicted on the basis of web pages features.
In link-based method, link-based ranking algorithms are used such as HITS and PageRank.

3.1 Machine learning method

One of the methods used to identify web spam is machine learning method. Since web spam methods are continuously changing, the classification of these methods should be necessarily temporary. However, there are some fixed principles [5]:

- Each successful spam, target one or more characteristics used by ranking algorithms of search engine.
- Web spam detection is a classification problem. Through using machine learning algorithms, search engines decide whether a page has spam or not.
- Generally, innovations in web spam detection are followed by statistical anomalies, and are related to some observable features in search engines.

Spam and nonspam pages have different statistical features [12], and these differences are used in terms of automatic classification. In this method at first, some features have been considered for spam page. Through using classification method and on the basis of these features, a method is learnt. On the basis of this method, search engine can classify pages into spam and nonspam page.

Ntoulas et al. took into account detection of web spam through content analysis [13]. Amitay et al. have considered categorization algorithms to detect the capabilities of a

website. They identified 31 clusters that each were a group of web spam [14].

Prieto et al. presented a system called SAAD in which web content is used to detect web spam. In this method, C4.5, Boosting and Bagging have been used for classification [15]. Karimpour et al. firstly reduced the number of features by using PCA, and then they considered semi-supervised classification method of EM-Naive Bayesian to detect web spam [16]. Rungsawang et al. applied ant colony algorithm to classify web spam. The results showed that this method, in comparison with SVM and decision tree, involves higher precision and lower Fall-out [17]. Silva et al. considered various methods of classification involving decision tree, SVM, KNN, LogitBoost, Bagging, adaBoost in their analysis[18]. Tian et al. have presented a method based on machine learning method, and used human ideas and comments and semi-supervised algorithm to detect web spam [19].

Becchetti et al. considered link based features such as TrustRank and PageRank to classify web spam [20]. Castillo et al. took into account link-based features and content analysis by using C4.5 classifier to classify web spam [21]. Dai et al. classified temporal features through using two levels of classification. The first level involves several SVM^{light}, and the second level involves a logistic regression [22].

3.2 Link-based method

With regard to emerging HITS and PageRank and the success of search engines in presenting optimized results by using link-based ranking algorithms, spammers tried to manipulate link structure to increase their own ranking.

PageRank method was introduced by Page et al. in 1998. This method was considered as one of the best solutions to combat with web spam. In this method, all links do not have the same weight in rank determination; instead, links from high rank sites present higher value in comparison with link of sites having fewer visitors. As a result, sites created by spammers rarely have a rule in determining the rank. Due to this issue, Google search engine has been preferred over years [23].

HITS method has been presented by Kleinberg. In this algorithm, sites are divided into two group; namely, Hubs and Authorities sites. In this algorithm, Hub sites refer to those sites involving many links in Authorities sites. These two group effect ranking [24]. Figure 1 show Hub and Authority sites.

Bharat and Henzinger presented imp algorithm proposed as HITS development to solve the problem of mutual reinforcement. Their idea is that if there is K edge on one site in the first host to one document in the second host, and then Authority weight is computed as 1/K. In contrast, if there is L edge from one document over the first host to a set of pages over the second host, then Hub weight is computed as 1/L [25].

Zhang et al. used the quality of both content and link to combat with web spam. They presented a repetitive procedure to distribute the quality of content and link in other pages of the web. The idea proposed in terms of combining content and link to detect link spam seems logical [26].

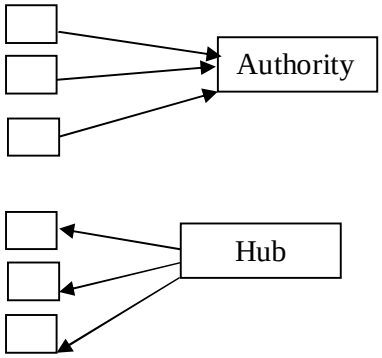

Figure 1. Hub and Authority

Acharya et al. proposed using historical data to detect spam pages for the first time. Heterogeneous growth rate in back links may be a signal of spam [27]. Also, shen et al. features extracted from various reports of web graph are growth rate in input link and death rate in input link.
Eiron et al. proposed HostRank that is more resistant against link spam in comparison with PageRank [28]. Lempel et al. proposed "TKC effect" for the first time. In this method, connected pages obtain high rank for iterative processes. Link farms misuse TKC effort to increase their own rank in search engines. They proposed SALSA algorithm that is more resistant against TKC effect in comparison with HITS [29].

Ng et al. proposed two algorithms; namely, random HITS and subspace THIS for the instability of HITS [30]. Zhang et al. proposed damping factors to compute PageRank to detect the collusion between web pages [31]. Li et al. presented some method to improve HITS results. According to HITS, these pages having less input links and more output link, undesirable results will be obtained. They proposed weighted setting for such pages in adjacently matrix to solve this problem [32]. Chakrabarti et al. created the model of DOM for each web page, and they found out that sub trees that correspond with searching more than other parts, show special behavior against the process of mutual reinforcement [33].

Gyngyi et al. used the concept of trust to combat with link spam, and proposed TrustRank algorithm. TrustRank is one of the most popular and successful anti-spamming techniques. TrustRank is based on trust concept in social networks. In this way, good pages usually point to good pages, and good pages rarely have links to spam pages. Therefore, at first, a group of valid pages are selected, and trust score is dedicated to them. Then, it is followed like distribution scheme of PageRank. Algorithm 1 shows TrustRank algorithm. This is not very different from computing main PageRank. In this algorithm, selecting the seed set is very important. Selection is performed in a way that those pages that have high PageRank score and connection are selected. Here, inverse PageRank is selected in order to select connected and seed pages.

Also, Gyngyi et al. presented different value of PageRank and TrustRank to precisely detect spam pages. In this way, the pages involving good PageRank score and weak TrustRank score are considered as link-based spam pages [34].

Input: T transition matrix
N number of pages
L limit of oracle invocations
α_B decay factor for biased PageRank
M_B number of biased PageRank iterations
Output:
t^* TrustRank scores
Begin
1 s ← SelectSeeds(. . .) ;
2 σ ← Rank({1, . . . , N}, s) ;
3 d ← 0_N ;
4 for i← 1 to L do
if O(σ (i)) = 1 then
d(σ (i)) ← 1 ;
5 d ← d/|d|;
6 t^* ← d ;
for i = 1 to M_B do
$t^* = \alpha_B \cdot T \cdot t^* + (1- \alpha_B)d$
return t^*
End

Algorithm 1. TrustRank

One of anti-spamming algorithms is BadRank. In this algorithm, bad initial page collection is selected, and a value is dedicated to each page in bad pages collection. In this algorithm, like PageRank, a bad value can be distributed via web graph repeatedly. In each repetition, bad value is dedicated to each page pointing to bad pages. Finally, spam pages will have bad and high scores[35].

Guha et al. proposed an algorithm of distributing trust and distrust values at one time [36]. Wu et al. as well as Krishnan and Raj proposed distrust distribution to combat with web spam [2,37]. Both results showed that using distrust distribution in reducing spam rank is more useful than using the trust alone.

Benczur et al. proposed SpamRank. According to their proposition, PageRank values of input link in normal pages should follow power rule distribution. They investigated PageRank distribution of all input links. If, a normal pattern is not followed by distribution, then a penalty will be considered for this page [38].

Becchetti et al. proposed Truncated PageRank algorithm to combat link-based spam. They suppose that link farm spam pages may involve many supporters in web graphs in short intervals, but they don't have any supporters in long intervals, or they have few supporters. Based on this assumption, they presented Truncated PageRank. The first level of links is ignored, and nodes of next stages are computed [39].

Another anti-spamming algorithm is "anti- TrustRank", and it is supposed that if a page points to bad pages, then it may be bad. This algorithm is inverted TrustRank. Anti-TrustRank distributes "bad" scores. In comparison with TrustRank, anti-TrustRank selects "bad" pages instead of good pages [37].

Spam Mass Estimation was introduced following TrustRank. Spam Mass is a measurement of how a page rank is created via linking by spam page. It computed and combines both scores involving regular and malicious scores [34].

Wu and Davison proposed Parent Penalty to combat with link farms [40]. Their algorithm involves three stages.

- Producing a seed set from all data collection
- Development stage
- Value ranking

Algorithm 2 shows that how initial collection is selected. Here, IN(P) shows a collection input links in page P. INdomain(P) and OUTdomain(P) show the domain of input links and output page of P respectively. d(i) is the name of link domain of i.

1 **for** p **do**

2 **for** i in IN(p) **do**

3 **if** d(i) ≠ d(p) and d(i) not in INdomain(p) **then** add d(i) to INdomain(i) ;

4 **for** k in IN(p) **do**

5 **if** d(k) ≠ d(p) and d(k) not in OUTdomain(p) **then** add d(k) to OUTdomain(i) ;

6 X ← the intersection of INdomain(p) and OUTdomain(p);

7 **if** size(X) ≥T_{IO} **then** A[p] ← 1 ;

Algorithm 2. ParentPenAlty: Seed Set

Pages in link farms usually have several nodes common between input and output links. If there is just one or two common nodes, then this page will not be marked as a problematic page. If there is more common nodes, then page may be a part of link farm. In this stage, T_{Io} threshold is used. When the number of common links of input and output links is equal to T_{Io} or greater than T_{Io}, page will be marked as spam, and it is placed in seed set.

Development stage has been shown in algorithm 3. In this stage, bad initial value is distributed for page. It is supposed that if a page only points to a spam page, then no penalty will be considered for it, while if a page involves many output links in spam pages, then the page may be a part of link farm. Hence, another threshold (T_{PP}) is used to detect a page. In this way, if the number of output links in spam pages is equal to threshold or more than threshold, then that page will be marked as spam.

Data: A[N], T_{PP}

1 **while** A do change **do**

2 **for** p : A[p] = 0 **do**

3 badnum ← 0 ;

4 **for** k ϵ OUT(p) **do if** A[k] = 1 **then** badnum ← badnum + 1 ;

5 **if** badnum ≥ T_{PP} **then** A[p] ← 1 ;

Algorithm 3. ParentPenalty: Seed Set Expansion

Finally, bad value is combined with normal link based ranking algorithms. In this way, adjacent matrix of web graph is changed in data set. There are two possibilities to consider a penalty for spam links. They are as follows: reducing the weight of adjacent matrix elements or removing link.

4. CONCLUSION

In the paper, web spam has been considered as a crucial challenge in the world of searching. We explained various methods of web spamming and algorithms to combat with web spam. Up to now, many methods have been created to combat with web spam. However, due to its economical profit and attractiveness, on one side, researchers have presented new methods to combat with it, and in another side, spammers present some methods to overcome these limitations. As a result, a certain method has not been proposed up to now. We hope that we can observe spam pages reduction by presenting character algorithms to detect web spams.

5. REFERENCES

[1] Gyongyi, Z. and H. Garcia-Molina, Web Spam Taxonomy, in First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005). 2005: Chiba, Japan.

[2] Wu, B., V. Goel, and B.D. Davison. Topical trustrank: Using topicality to combat web spam. in Proceedings of the 15th international conference on World Wide Web. 2006. ACM.

[3] Gyngyi, Z. and H. Garcia-Molina, Link spam alliances, in Proceedings of the 31st international conference on Very large data bases. 2005, VLDB Endowment: Trondheim, Norway. p. 517-528.

[4] Abernethy, J., O. Chapelle, and C. Castillo, WITCH: A New Approach to Web Spam Detection, in In Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb}. 2008.

[5] Najork, M., Web Spam Detection. Encyclopedia of Database Systems, 2009. **1**: p. 3520-3523.

[6] Castillo, C., et al., A reference collection for web spam. SIGIR Forum, 2006. **40**(2): p. 11-24.

[7] Davison, B.D., Recognizing nepotistic links on the web. Artificial Intelligence for Web Search, 2000: p. 23-28.

[8] Collins, G. Latest search engine spam techniques. Aug 2004; Available from: http://www.sitepoint.com/article/search-engine-spam-techniques.

[9] Perkins, A. The classification of search engine spam. 2001; Available from: http://www.silverdisc.co.uk/articles/spam-classification.

[10] Sasikala, S. and S.K. Jayanthi. Hyperlink Structure Attribute Analysis for Detecting Link Spamdexing. in International Conference on Advances in Computer Science–(AET-ACS 2010), Kerela. 2010.

[11] Wu, B. and B.D. Davison. Cloaking and Redirection: A Preliminary Study. in AIRWeb. 2005.

[12] Fetterly, D., M. Manasse, and M. Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. in Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004. 2004. ACM.

[13] Ntoulas, A., et al. Detecting spam web pages through content analysis. in the 15th International World Wide Web Conference. May 2006. Edinburgh, Scotland.

[14] Amitay, E., et al. The connectivity sonar: Detecting site functionality by structural patterns. in the 14th ACM Conference on Hypertext and Hypermedia. Aug 2003. Nottingham, UK.

[15] Prieto, V., et al., Analysis and Detection of Web Spam by Means of Web Content, in Multidisciplinary Information Retrieval, M. Salampasis and B. Larsen, Editors. 2012, Springer Berlin Heidelberg. p. 43-57.

[16] Karimpour, J., A. Noroozi, and S. Alizadeh, Web Spam Detection by Learning from Small Labeled Samples. International Journal of Computer Applications, 2012. **50**(21): p. 1-5.

[17] Rungsawang, A., A. Taweesiriwate, and B. Manaskasemsak, Spam Host Detection Using Ant Colony Optimization, in IT Convergence and Services, J.J. Park, et al., Editors. 2011, Springer Netherlands. p. 13-21.

[18] Silva, R.M., A. Yamakami, and T.A. Alimeida. An Analysis of Machine Learning Methods for Spam Host Detection. in 11th International Conference on Machine Learning and Applications (ICMLA). 2012.

[19] Tian, Y., G.M. Weiss, and Q. Ma. A semi-supervised approach for web spam detection using combinatorial feature-fusion. in GRAPH LABELLING WORKSHOP AND WEB SPAM CHALLENGE. 2007.

[20] Becchetti, L., et al. Link-Based Characterization and Detection of Web Spam. in AIRWeb 2006. 2006. Seattle, Washington, USA.

[21] Castillo, C., et al., Know your neighbors: web spam detection using the web topology, in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007, ACM: Amsterdam, The Netherlands. p. 423-430.

[22] Dai, N., B.D. Davison, and X. Qi, Looking into the past to better classify web spam, in Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web. 2009, ACM: Madrid, Spain. p. 1-8.

[23] Page, L., et al., The PageRank citation ranking: bringing order to the web. 1999.

[24] Kleinberg, J.M., Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 1999. **46**(5): p. 604-632.

[25] Bharat, K. and M.R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. 1998. ACM.

[26] Zhang, L., et al. Exploring both content and link quality for anti-spamming. in Computer and Information Technology, 2006. CIT'06. The Sixth IEEE International Conference on. 2006. IEEE.

[27] Acharya, A., et al., Information retrieval based on historical data. 2008, Google Patents.

[28] Eiron, N., K.S. McCurley, and J.A. Tomlin, Ranking the web frontier, in Proceedings of the 13th

international conference on World Wide Web. 2004, ACM: New York, NY, USA. p. 309-318.

[29] Lempel, R. and S. Moran, The stochastic approach for link-structure analysis (SALSA) and the TKC effect. Computer Networks, 2000. **33**(1): p. 387-401.

[30] Ng, A.Y., A.X. Zheng, and M.I. Jordan. Stable algorithms for link analysis. in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001. ACM.

[31] Zhang, H., et al., Making eigenvector-based reputation systems robust to collusion, in Algorithms and Models for the Web-Graph. 2004, Springer. p. 92-104.

[32] Li, L., Y. Shang, and W. Zhang. Improvement of HITS-based algorithms on web documents. in Proceedings of the 11th international conference on World Wide Web. 2002. ACM.

[33] Chakrabarti, S., M. Joshi, and V. Tawde, Enhanced topic distillation using text, markup tags, and hyperlinks, in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001, ACM: New Orleans, Louisiana, USA. p. 208-216.

[34] Gyongyi, Z., et al., Link spam detection based on mass estimation, in Proceedings of the 32nd international conference on Very large data bases. 2006, VLDB Endowment: Seoul, Korea. p. 439-450.

[35] Sobek, M., Pr0-google's pagerank 0 penalty. badrank. 2002.

[36] Guha, R., et al., Propagation of trust and distrust, in Proceedings of the 13th international conference on World Wide Web. 2004, ACM: New York, NY, USA. p. 403-412.

[37] Krishnan, V. and R. Raj. Web spam detection with anti-trust rank. in the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2006). 2006. Seattle, USA.

[38] Benczur, A.A., et al. SpamRank–Fully Automatic Link Spam Detection Work in progress. in Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web. 2005.

[39] Becchetti, L., et al. Using rank propagation and probabilistic counting for link-based spam detection. in Proc. of WebKDD. 2006.

[40] Wu, B. and B.D. Davison. Identifying link farm spam pages. in Special interest tracks and posters of the 14th international conference on World Wide Web. 2005. ACM.