International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 34 - 48, 2016, ISSN:- 2319–8656

Efficient Fuzzy-Based System for the Diagnosis and Treatment of Tuberculosis (EFBSDTTB)

Angbera, Ature
Dept. Of Mathematics/ Statistics/Computer Science
University of Agriculture Makurdi
Makurdi, Nigeria

Esiefarienrhe, Michael Bukohwo
Dept. Of Mathematics/ Statistics/Computer Science
University of Agriculture Makurdi
Makurdi, Nigeria

Agaji, Iorshase
Dept. Of Mathematics/ Statistics/Computer Science
University of Agriculture Makurdi
Makurdi, Nigeria

Abstract: The aim of this study is to design a FuzzyBased Expert System for Tuberculosis diagnosis and Treatment. The designed system made use of General Hospital Adikpo, patient database. The system has 18 input fields and five outputs field. Input fields are Chest pain (CP), cough duration (CD), fever duration (FV), night sweats (NS), weight loss (WL), loss of appetite (LOA), change in bowel habits (CBH), variations in mental behaviour (VMB), masses along the neck (MAN), draining sinus (DS), coma (seizure) (CO), stiff Neck (SN), headache (HD), abdominal Pain (AP), painful or uncomfortable urination (PU), hemopysis (coughing up blood) (CUB), fatigue (FA) and blood present in urine (BPU). The output fields refers to the class/group of tuberculosis disease in the patient. This system uses Mamdani inference method. The results obtained from designed system are compared with the data in the database and observed results of designed system are correct. The system was designed with Java (Jfuzzylogic), Microsoft visio (2013), mySql workbench, MySql database, JSP and XHML.

Keywords: Expert System, Diagnosis, Treatment, Inference component, Fuzzy Logic, Membership Function and Rules Blocks

1. INTRODUCTION

Computer technology tools help doctors to organize, store and retrieve relevant medical knowledge needed to understand the problematic cases and give them ideas about a proper diagnosis, prognosis and treatment decisions. There are huge data management tools available within health care systems, but analysis tools are not sufficient to discover hidden relationships amongst the data [9]. Expert Systems (ES) of an intelligent computer is based on interactive decision tool that uses facts and rules to solve real life problems, based on knowledge obtained from one or more of a human expert in a specific area.

In domain of disease like tuberculosis, which is one of the killer disease in developing countries. This disease has the following symptoms: fever, chest pain, coughing up blood, stiff neck, abdominal pain, variation in metal behavior, night sweat, urinating blood etc. Because of the many and uncertain risk factors in tuberculosis disease, sometimes the disease diagnosis is hard for experts.

Having so many symptoms to analyze to diagnose tuberculosis of a patient, the physician job is made very difficult. So, experts require an accurate tool that considering these numerous

symptoms, a system that can diagnose and prescribe treatment (drugs) for tuberculosis in our health care centers is very important. Motivated by the need of such an important tool, in this study, we designed an expert system to diagnose tuberculosis disease and prescribe drugs for the patient. The designed expert system is based on Fuzzy Logic.

Fuzzy Logic is a form of multi-valued logic derived from fuzzy set theory to deal with approximate reasoning. It provides the means to represent and process the linguistic information and subjective attributes of the real world [4]. Most of the systems that are constructed based on fuzzy sets and logic have a common architecture. This fuzzy logic systems are based on a specific lifecycle model consisting of four characteristic stages namely, Fuzzifier component, Fuzzy Inference Engine, Fuzzy Rules and Defuzzifier component [15].

2. REVIEW OF LITERATURES

Tuberculosis (TB) is an infectious disease caused by mycobacteria, mainly *Mycobacterium tuberculosis*. It commonly attacks the lungs (pulmonary TB) but can also affect the central nervous system, the lymphatic system, the circulatory system, the genitourinary system, bones, joints and even the skin. Other mycobacteria such as *Mycobacterium bovis*, *Mycobacterium africanum*, *Mycobacterium confetti* and *Mycobacterium microti*, can also cause tuberculosis, but these species do not usually infect healthy adults, [11].

[5], proposed a fuzzy expert system for tuberculosis diagnosis which was developed for providing decision support platform to tuberculosis researchers, physicians, and other healthcare practitioners in tropical medicine. The combination of inadequate expertise and sometimes the complexity of medical practices exponentially increase the morbidity and mortality rates of tuberculosis patients. The task of arriving at an accurate medical diagnosis may sometimes become very complex and cumbersome. Fuzzy logic technology provides a simple way to arrive at a definite conclusion from vague, imprecise and ambiguous medical data. In order to achieve this, a study of the knowledge base system for tuberculosis was undertaken and a fuzzy expert system was developed using fuzzy logic technology [5]. Their system composed of four components which include the knowledge base, the fuzzification, the inference engine and defuzzification components. The fuzzy inference method employed in the research was the Root Sum Square (RSS). Triangular membership function was used to show the degree of participation of each input parameter and the defuzzification technique employed in this research is the Center of Gravity (CoG). The fuzzy expert system was designed based on clinical observations, medical diagnosis and the expert's knowledge. They selected 30 patients with tuberculosis and computed the results that were in the range of predefined limits by the domain experts [5]. [13], proposed a rule based Fuzzy Diagnostics and a decision support system which was intended to be used by pulmonary physicians, which will analyze the class of tuberculosis, by providing inputs (TB symptoms) into the system. In the formulation of fuzzy set system, the ranges of scores were classified in each symptom. The values or scores had undergone the process of fuzzification, which was also responsible for the threshold calculations that are needed by system for some reasoning, which are included in fuzzy relations. After processing the calculations, the resultant scores were graphed in a symmetrical manner. The graph will illustrate the scores and its corresponding membership values. After the graph process, the fuzzy logic sets were intersected and it determined a matrix format. The matrix illustrated that the symptoms are between the intersection points. The rules were determined by the scores that had undergone defuzzification process.

[1], designed an expert system for diseases diagnosis using Fuzzy set. In their approach, they used fuzzy set to diseases diagnosis, this was depending on opinion of 20 doctors. The result of the system shows the diagnosis of three types of

respiratory diseases, which tuberculosis is one of them. The system used four symptoms, namely: X- ray, Respiratory rate (RR), Cough (CO) and Fever (F) which were indicated as input of the fuzzy logic and the output was in a range of the risks and type of respiratory diseases.

[7], developed a diagnostic fuzzy cluster means system to help in diagnosis of Tuberculosis using a set of symptoms. The system which uses a set of clustered data set was more precise than the traditional system. The classification, verification and matching of symptoms to the seven groups of clusters was necessary especially in some complex scenarios. The model proposed allows for the classification and matching of cluster groups to TB symptoms.

[14], proposed the fuzzy Artificial Immune Recognition System (AIRS), for tuberculosis diagnosis detection. In designing the system, the Fuzzy Logic Controller was applied, which converts the continuous inputs into fuzzy sets. Ten features of tuberculosis diagnosis were defined for the fuzzy input. Asthma is a chronic inflammatory lung disease. An automated system was developed using a self-organizing fuzzy rule-based system [2].

According to [12], Arthritis is a chronic disease and about three fourth of the patients are suffering from osteoarthritis and rheumatoid arthritis which are undiagnosed and the delay of detection may cause the severity of the disease at higher risk. A system for the diagnosis of Arthritis using fuzzy logic controller (FLC) was designed which was, a successful application of Zadeh's fuzzy set theory [16]. It is a potential tool for dealing with uncertainty and imprecision [12]. [10], designed a decision support system for malaria and dengue (DSSMD). The diagnosis of disease was solely based on the non - clinical symptoms of the disease using Artificial intelligence. In another study [6] developed a fuzzy expert system for the management of *malaria* (FESMM) which provides decision support platform to malaria researchers, physicians to assist malaria researchers, physicians and other health practitioners in malaria endemic regions.

2.1 TB Disease Treatment Regimens

In a study by [3], there are four basic treatment regimens recommended for treating patients with TB disease caused by organisms that are known or presumed to be susceptible to Isoniazid (INH), Rifampin (RIF), Pyrazinamide (PZA), and Ethambutol (EMB). Each treatment regimen consists of an initial 2-month treatment phase followed by a continuation phase of either 4 or 7 months. The 4-month continuation phase is used for the majority of patients. Although these regimens are broadly applicable, there are modifications that should be made under specified circumstances.

3. METHODOLOGY

In this section, we show the fuzzy expert system designed, membership functions fuzzification, fuzzy rule base and defuzzification.

3.1 The Proposed System

The system is designed to aid in the diagnosis and treatment of TB in our settings. The success of any Fuzzy Expert System depends upon the opinion of the domain experts on various issues related to the field of study. The developed Fuzzy Expert System for the Management of tuberculosis has an architecture as shown in Figure 1 below.

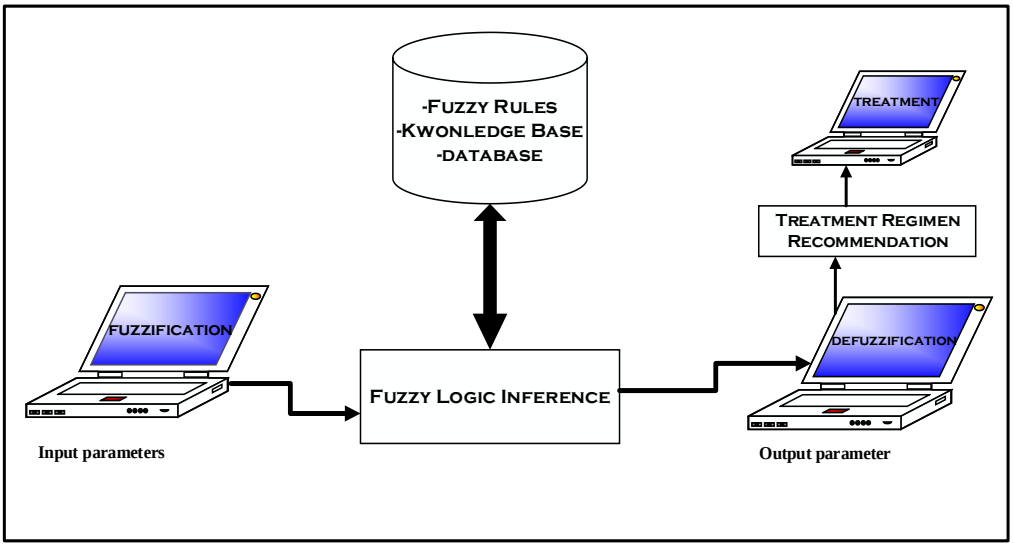


Figure 1: Architectural Design of the Proposed Diagnosis and Treatment of TB.

3.2 Inputs/Output Membership Function (Fuzzification/Defuzzification) for the Proposed System.

In this section we defined fuzzy input parameters which are the symptoms of TB with their linguistic categories described as *low*, *medium* and *high* to be used for the diagnosis and treatment based on the classes of TB. The fuzzy input parameters (symptoms scores) we be fuzzified using triangular membership functions. The inputs for the proposed system are: Chest pain (CP), cough duration (CD), fever duration (FV), night sweats (NS), weight loss (WL), loss of appetite (LOA), change in bowel habits (CBH), variations in mental behavior (VMB), masses along the neck (MAN), draining sinus (DS), coma (seizure) (CO), stiff Neck (SN), headache (HD), abdominal Pain (AP), painful or uncomfortable urination (PU), hemopysis (coughing up blood) (CUB), fatigue (FA) and blood present in urine (BPU). This inputs parameters are model in a range of six weeks. The low level of the symptoms shows that TB is just beginning in the patient body, the medium stage indicate that the patient situation is becoming bad and the final stage which is the high level indicate that the patient case has gone to a very bad state, which if not carefully handled the patient might loss his/her life. This scenario is the same for all the TB symptoms to consider in this research work .Figures 3.2a and 3.2b shows the membership function distribution of the inputs symptoms.

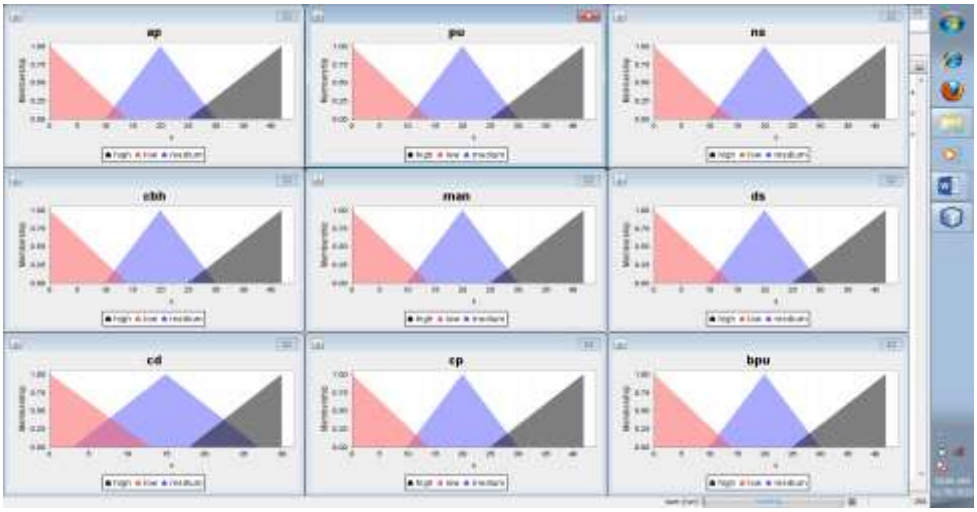


Figure 2a: membership function of the inputs variables

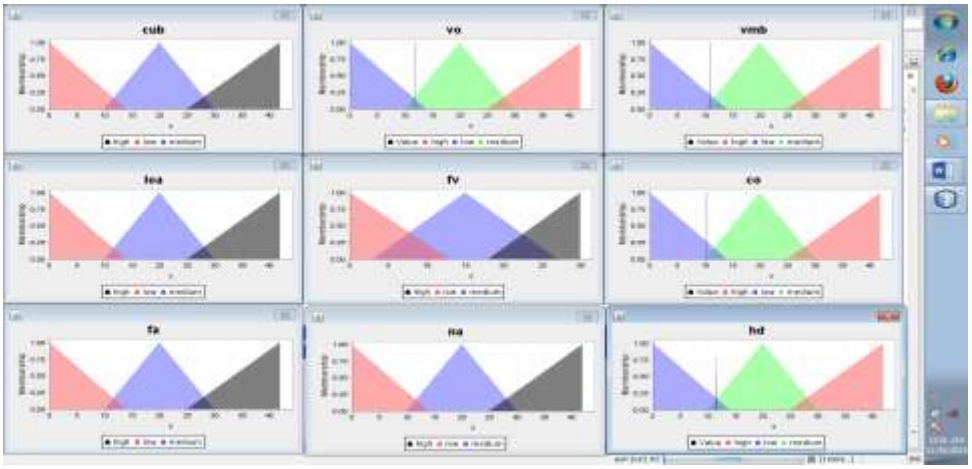


Figure 2b: membership function of the inputs variable

3.4 The Diagnosis/Output Design for System

This subsection describes the output of the fuzzy logic engine. There are five output variables, we shall call them "Tuberculosis Group", namely: Tuberculosis meningitis (TBM), Gastrointestinal tuberculosis (GITB), Tuberculosis lymphadenitis (TBL), Genitourinary tuberculosis (GUTB) and

Pulmonary tuberculosis (PTB). Here too the triangular membership functions will be used on each output variable. The membership functions details are shown in Table 1.

Table 1 The Output variable (Tuberculosis Group) ranges that correspond to each fuzzy set

Output field	**Range**	**Fuzzy Set**
TUBERCULOSIS GROUP (TBG)	0<TBG<3	TBM1
	3<TBG<7	TBM2
	7<TBG<10	TBM3
	0<TBG<3	GITB1
	3<TBG<7	GITB2
	7<TBG<10	GITB3
	0<TBG<3	TBL1
	3<TBG7	TBL2
	7<TBG<10	TBL3
	0<TBG<3	GUTB1
	3<TBG<7	GUTB2
	7<TBG<10	GUTB3
	0<TBG<3	PTB1
	3<TBG<7	PTB2
	7<TBG<10	PTB3

From table 1 it shows that, the higher the value, the higher the health risk of the patient. In this system, we have 15 fuzzy sets for the output variable risk group (TBM1, TBM2, TBM3, GITB1, GITB2, GITB3, TBL1, TBL2 TBL3, GUTB1, GUTB2, GUTB3, PTB1, PTB2 and PTB3). Each of the five output variable has three fuzzy set, which represent the low, medium and high level of TB disease to be diagnosed. To get a particular class of TB, a set of inputs (symptoms) must be inputted into the system. This inputs will also identify the degree of risk of the patient whether the level of TB is low, medium or high. For example, to detect T BM1, TBM2 or TBM3. if the values of inputs are in the range of 0 to 3, then the level of TB is TBM1 which is low, also if the values of inputs are in the range of 3 to 7, then the TB level is TBM2 which is medium and if the values inputted are in the range of 7 to 10, then then level of TB is TBM3, which is high. The mode of evaluation is the same to other classes of TB to be diagnosed by the system. Figure 3 shows the membership functions of the output variables.

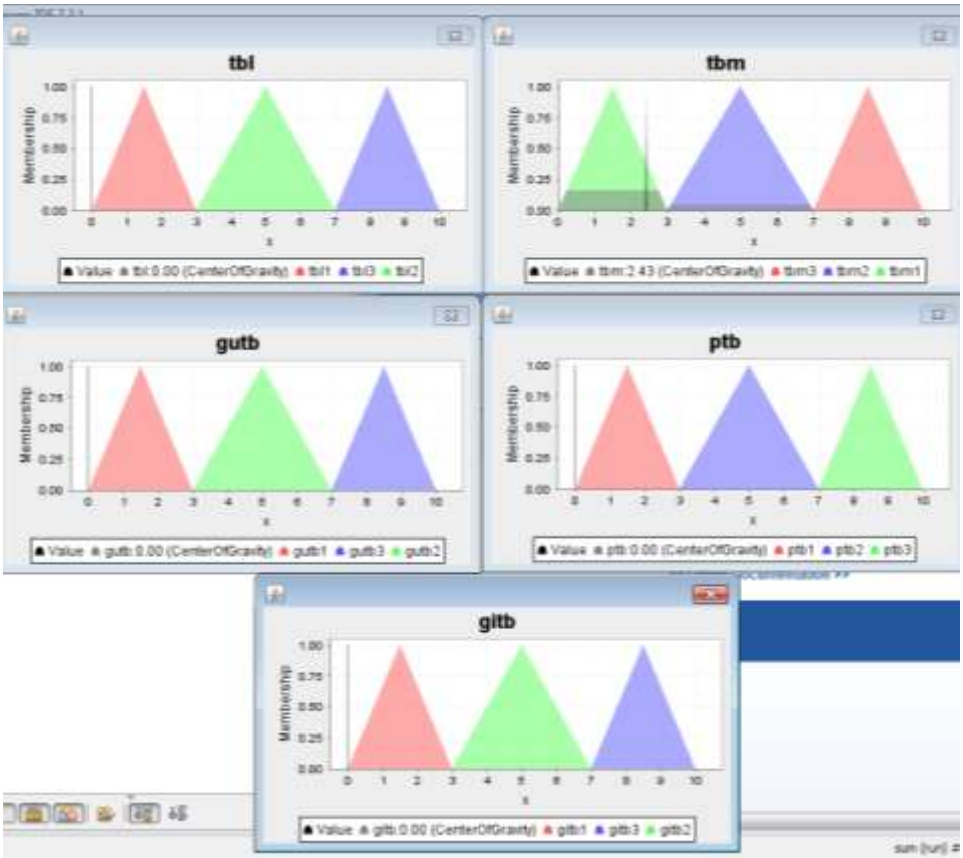


Figure 3: membership functions of the output (TBG) variables.

3.5 Activity Diagram

Activity diagram describes the business and operational step-by-step workflows of components in a system. It shows the overall flow of control detailing the sequence of activities from a start point to the finish point displaying the many decision paths that exist in the progression of events contained in the activity. They may be used to detail situations where parallel processing may occur in the execution of some activities. Activity diagram for the fuzzy-based system is shown in figure 4. Before the doctor accesses the system, he/she must be authenticated.

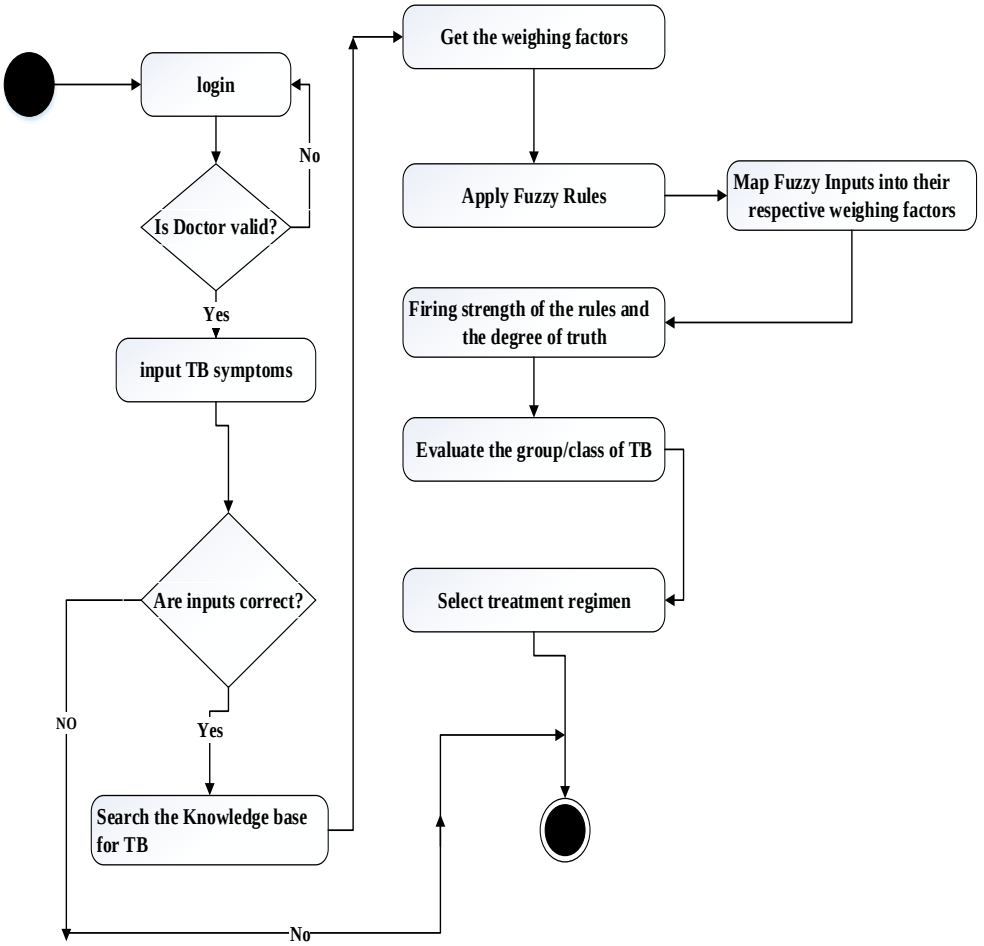


Figure 4: Activity Diagram for the Proposed Diagnosis and Treatment of TB.

3.6 Defuzzification of the Outputs for the System

The input for the defuzzification process is a fuzzy set (the aggregate output fuzzy set), and the output is a single number. As much as fuzziness helps the rule evaluation during the intermediate steps, the final desired output for each variable is generally a single number. However, the aggregate of a fuzzy set encompasses a range of output values, and so they must be defuzzified in order to resolve a single output value from the set. Perhaps the most popular defuzzification method is the centroid or fuzzy centroid, this method returns the centre of an area under the curve. . It computes the defuzzified result by determining the element for which the combined set's area is divided into two equal faces (Klir and Yuan, 1995). This is the method adopted. The defuzzification of the data into a crisp output is accomplished using the "Fuzzy Centroid".

Fuzzy Centroid Formula is as shown in equation 1.

$$Output = \frac{\sum_{i=1}^{n}(Center_i * Strength_i)}{\sum_{i=1}^{n} Strength_i} \quad 1$$

Where n, is the number of output members.

3.7 Fuzzy Inference Component for the System

A fuzzy inference system is usually composed of one or more Function Blocks (FB). Each FB has variables (input, output or instance variables) as well as one or more Rule Blocks (RB). Each rule block is composed of a set of rules, as well as Aggregation, Activation and Accumulation methods. All methods defined in the norm are implemented in jFuzzyLogic. We will adhere to the definitions of Aggregation, Activation and Accumulation as defined by IEC-61131-7. Aggregation methods define the t-norms and t-conorms playing the role of intersection, union and complement operators. Activation method define how rule antecedents modify rule consequents, i.e. once the IF part has been evaluated, how this result is applied to the THEN part of the rule. Some of the fuzzy rules for this work are as shown below.

RULEBLOCK No1

AND: MIN; // Use 'min' for 'and' (also implicit use

//'max' for 'or' to fulfill DeMorgan's Law)

ACT: MIN; // Use 'min' activation method

ACCU: MAX; // Use 'max' accumulation method

RULE 1: IF sn IS low AND hd IS low AND vmb IS low AND co IS low AND vo IS low THEN tbm IS tbm1;

RULE 2: IF sn IS high AND hd IS high AND vmb IS high AND co IS high AND vo IS high THEN tbm IS tbm3;

RULE 3: IF sn IS medium AND hd IS medium AND vmb IS medium AND co IS medium AND vo IS medium THEN tbm IS tbm2;

RULE 4: IF ap IS low AND fv IS low AND wl IS low AND cbh IS low AND vo IS low AND na IS low THEN gitb IS gitb1;

RULE 5: IF ap IS medium AND fv IS medium AND wl IS medium AND cbh IS medium AND vo IS medium AND na IS medium THEN gitb IS gitb2;

RULE 6: IF ap IS high AND fv IS high AND wl IS high AND cbh IS high AND vo IS high AND na IS high THEN gitb IS gitb3;

RULE 7: IF man IS low AND ds IS low THEN tbl IS tbl1;

RULE 8: IF man IS medium AND ds IS medium THEN tbl IS tbl2;

RULE 17: IF man IS high AND ds IS low THEN tbl IS tbl2;

RULE 97: IF man IS medium AND ds IS high THEN tbl IS tbl2;

RULE 98: IF man IS high AND ds IS medium THEN tbl IS tbl2;

RULE 99: IF man IS medium AND ds IS low THEN tbl IS tbl1;

RULE 100: IF man IS low AND ds IS medium THEN tbl IS tbl1;

END_RULEBLOCK

END_FUNCTION_BLOCK

3.8 Treatment Regimen for TB

In the treatment of TB there are four basic treatment regimens recommended for treating patients with TB disease caused by organisms that are known. These drugs include Isoniazid (INH), Rifampin (RIF), Pyrazinamide (PZA), and Ethambutol (EMB), etc. Each treatment regimen consists of an initial 2-month treatment phase followed by a continuation phase of either 4 or 7 months. The 4-month continuation phase is used for the majority of patients.

4. RESULTS

The welcome/Login interface for the designed Fuzzy-based system for TB diagnosis is as shown in figure 5

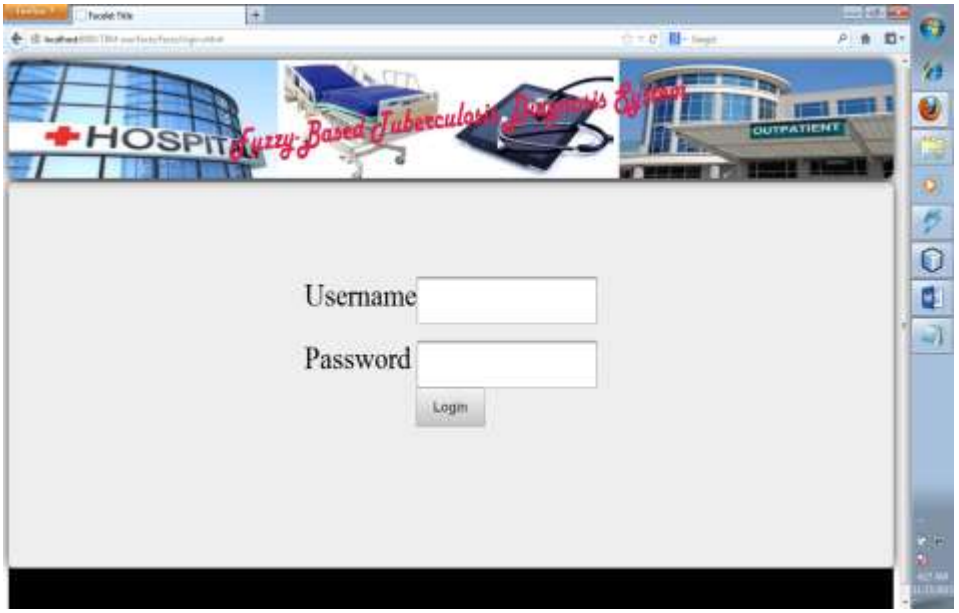


Figure 5: Login interface/welcome page.

Symptom selection interface for the designed Fuzzy-based system for TB diagnosis is as shown in figure 6

Figure 6: symptoms selection interface

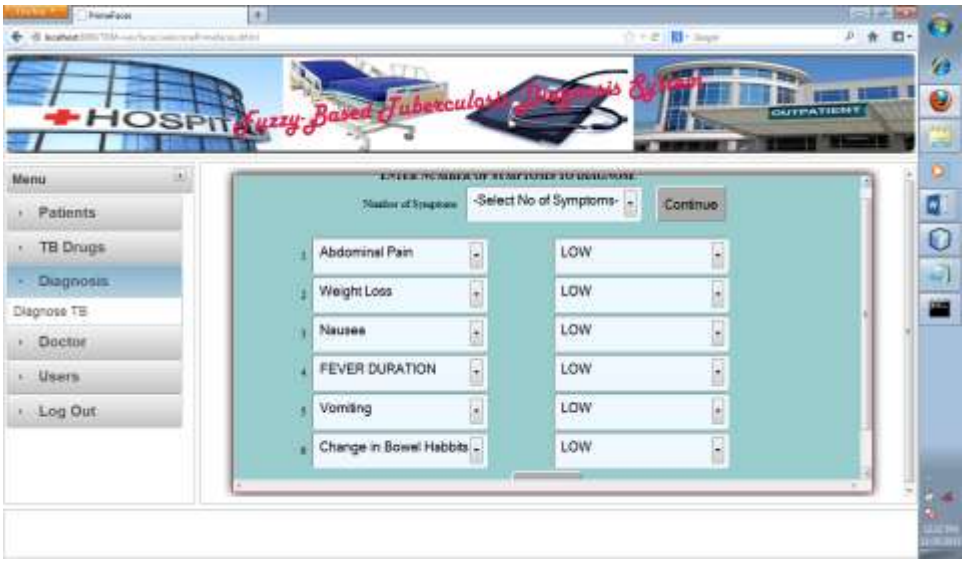


Figure 7: Symptoms Selection for a particular Case.

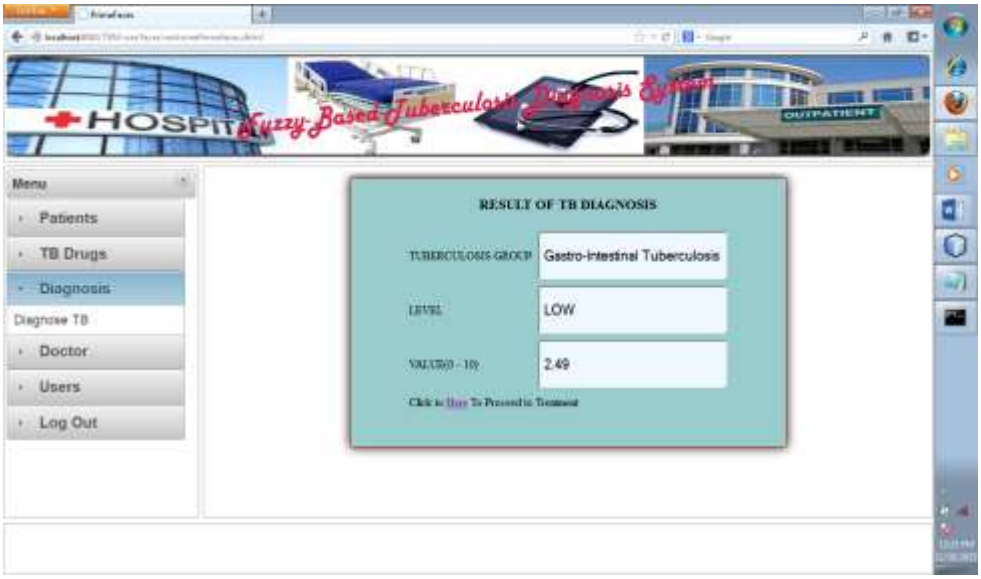


Figure 8: Diagnosed Result for Gastrointestinal TB

Figure 9: Five (5) Symptoms Selected for a particular Case.

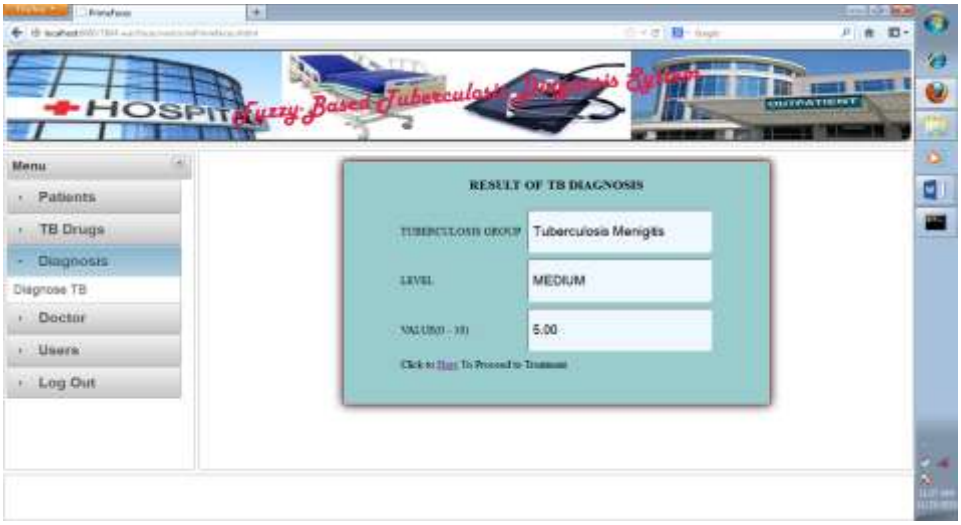


Figure 10: Diagnosed Result for Tuberculosis Menigitis

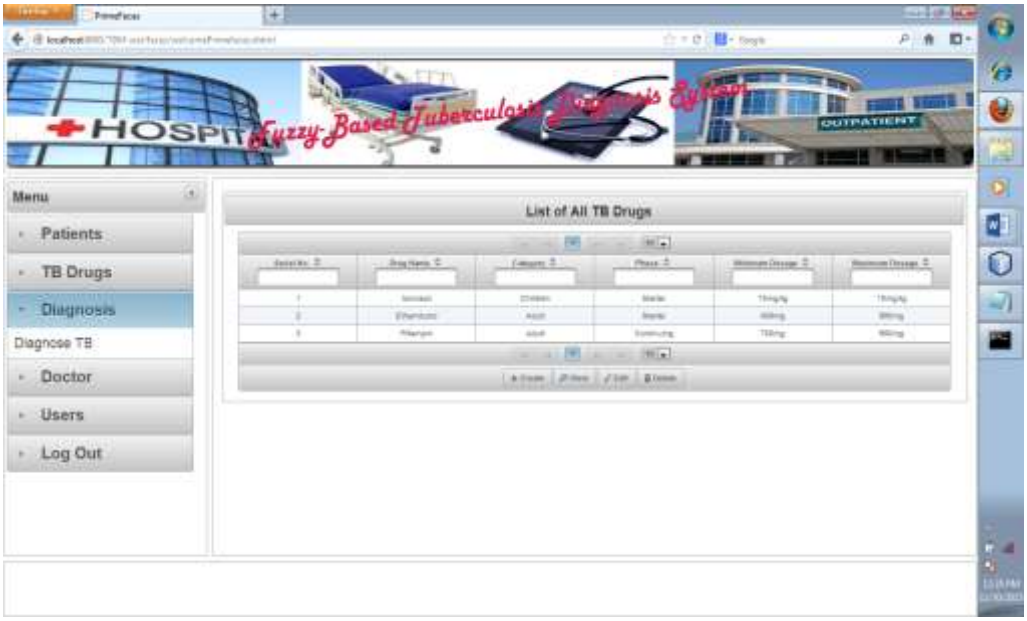


Figure 11: Drugs list Selections for Treatment

5. DISCUSSIONS

From figure 5, a user must be authenticated before he/she is allowed to use the system, on gaining access into the system, the user (health personnel) will select the symptoms selections interface as shown in figure 6, to fill in the

symptoms as complained by the patient. Figure 7 shows a patient symptoms were selected and their intensities (abdominal pain was low, weight loss was low, nausea was low, fever duration was low, vomiting was low and change in bowel habit was low) upon diagnosis, the result was found to be Gastrointestinal TB with a degree risk value of 2.49 which is low and agreed with our design from table 1 (0 < TBG <3). From figure 9 also another patient complained were selected with their degree of intensities (stiff neck was low, headache was medium, vomiting was medium, variation in metal behavior was medium and coma was medium) upon diagnosis the result was Tuberculosis Menigitis with a degree risk value of 5.00 as shown in figure 10, which also agreed with our design from table 1 (3 < TBG < 7). Figure 11 shows the list of drugs to be selected for treatment if the patient is affected with any of the TBG as designed in this research.

6. CONCLUSION

This Fuzzy-Based System for Tuberculosis Diagnosis and Treatment was designed with Java (JFuzzy Logic), membership functions, input variables, output variables and rule base. In this research there are 18 inputs or input variables and five outputs or output variables. The designed system has been tested with expert-doctor. This system is one of the simple and more efficient method for the diagnosis of TB, within a short possible time.

For further work, we recommend that other diseases like cancer, be integrate into the system and also make it a mobile application.

7. REFERENCES

[1] Abbas K., XuDe Z and Shaker K. (2010), Novel Respiratory Diseases Diagnosis by Using Fuzzy Logic. *Global Journal of Computer Science and Technology* Vol. 10 Issue 13 (Ver. 1.0).

[2] Ashish P, Jyotsna C, Shailendra K, Gupta M, VermaMembe K., Rajendra P, Qamar R, (2012), " Decision Support System for the Diagnosis of Asthma Severity Using Fuzzy Logic". Proceedings of the International MultiConference of Engineers and Computer Scientists 2012 Vol I, IMECS 2012, march 14-16, Hong Kong

[3] CDC (2013), Core Curriculum on Tuberculosis: What the Clinician Should Know. Centers for Disease Control and Prevention National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention Division of Tuberculosis Elimination. Retrieved from www.cdc.gov/tb.

[4] Chuen C. (1990), Fuzzy logic control system: Fuzzy logic controller –part I —, IEEE *Transaction on systems, man, and cybernetics*, Vol.20, No.2, pp.404~418.

[5] Djam, X.Y. and Y.H. Kimbi. (2011), A Decision Support System for Tuberculosis Diagnosis. *Pacific Journal of Science and Technology*. 12(2):410-425.

[6] Djam X.Y., G. M. Wajiga, Y. H. Kimbi and N.V. Blamah, (2012) "A Fuzzy Expert System for the Management of Malaria",International Journal of Pure and Applied Sciences and Technology.

[7] Imianvan A.A. and Obi J.C. (2011), Fuzzy Cluster Means Expert System for the Diagnosis of Tuberculosis. *Global Journal of Computer Science & Technology* Volume 11 Issue Version 1.0.

[8] Klir G. and Yuan B. (1995). Fuzzy Sets and Fuzzy Logic: Theory and Applications journal, *4*(1), pp1-4.

[9] Kumar S and Kaur G (2013), Detection of Heart Diseases using Fuzzy Logic. *International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 6.*

[10] Priynka S, Singh D, Manoj K and Nidhi M (2013), Decision Support System for Malaria and Dengue Disease Diagnosis (DSSMD*). International Journal of Information and Technology.* Volume 3, Number 7, pp. 633-640. http://www. irphouse.com /ijict.htm.

[11] Raviglione, M. C. and O' Brien, R. J. (2004). Tuberculosis in Kasper DL., Braunwald E, Fauci As, Hauser SL, Longo DL Jameson JL, Isselbacher KJ, eds.: Hanison's Principles of Internal Medicine,

16th ed., Mc Graw – Hill Professional, 953 – 66.

[12] Singh S, Kumar A, Panneerselvam K and Vennila J. (2012), Diagnosis of arthritis through fuzzy inference system. *Journal of Medical systems* 2012 June

[13] Soundararajan K, Sureshkumar S and Anusuya C (2012), Diagnostics Decision Support System for Tuberculosis using Fuzzy Logic. *International Journal of Computer Science and Information Technology & Security (IJCSITS)*, Vol. 2, No.3.

[14] Shamshirband S, Somayeh H, Hossein J, Mohsen A, Shaghayegh V, Dalibor P, Abdullah G and Kiah L (2014), Tuberculosis Disease Diagnosis Using Artificial Immune Recognition System. *International Journal of Medical Sciences* 11(5): 508-514.

[15] William G. (2011). An Optimization Approach to Employee Scheduling Using Fuzzy Logic. (MSc. Thesis, California Polytechnic State University, San Luis Obispo).

[16] Zadeh L (1965), "Fuzzy Sets," Information and Control, Vol. 8, No. 3.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 49 - 55, 2016, ISSN:- 2319–8656

An Effective Approach for Document Crawling With Usage Pattern and Image Based Crawling

Ankur Tailang
School Of Information Technology
Mats University
Raipur,India

Abstract: As the Web continues to grow day by day each and every second a new page gets uploaded into the web; it has become a difficult task for a user to search for the relevant and necessary information using traditional retrieval approaches. The amount of information has increased in World Wide Web, it has become difficult to get access to desired information on Web; therefore it has become a necessity to use Information retrieval tools like Search Engines to search for desired information on the Internet or Web. Already Existing and used Crawling, Indexing and Page Ranking techniques that are used by the underlying Search Engines before the result gets generated, the result sets that are returned by the engine lack in accuracy, efficiency and preciseness. The return set of result does not really satisfy the request of the user and results in frustration on the user's side. A Large number of irrelevant links/pages get fetched, unwanted information, topic drift, and load on servers are some of the other issues that need to be caught and rectified towards developing an efficient and a smart search engine. The main objective of this paper is to propose or present a solution for the improvement of the existing crawling methodology that makes an attempt to reduce the amount of load on server by taking advantage of computational software processes known as "Migrating Agents" for downloading the related pages that are relevant to a particular topic only. The downloaded Pages are then provided a unique positive number i.e. called the page has been ranked, taking into consideration the combinational words that are synonyms and other related words, user preferences using domain profiles and the interested field of a particular user and past knowledge of relevance of a web page that is average amount of time spent by users. A solution is also been given in context to Image based web Crawling associating the Digital Image Processing technique with Crawling.

Keywords: WebCrawler, Page Ranking, Indexer, Usage Pattern, Relevant Search, Domain Profile, Migrating Agent, Image Based Crawling.

1. INTRODUCTION:

World Wide Web Is the largest hub for getting data related to any field. From the past few years it has become the major and the biggest means of getting information. Each n every day millions of pages get uploaded in the web related to some field, adding to the humongous number of millions pages already on-line. [1]As the rapid growth of World Wide Web from past ten years, it becomes difficult to get the desired information which user wants. The relevancy guaranteed by search engine is lack in accuracy. Search engine have some issue that need to be addressed to make it more efficient for the user, so that they can have more relevant page according to their previous requests. "This issue of search engine is like big number of irrelevant or unwanted links, topic drift and l server load" [6] that causes server failure. As with the help of search engine user query for their desired information, they generally entered some query with specific keywords what they wishes to access and search engine returns the list of URL's that are related to user keyword. Page rank ranks all the web pages according to their ranking to present in straighten out manner. Search engine may suffer many difficulties like
sometimes crawler download the irrelevant links and due to this quality of search engine reduces. To overcome this multiple crawling instances is introduced but it may results in network congestion also put extra burden on server.
Many algorithm like Page ranking [2] and HITS [3] etc. are used for ranking. But there is no contingency given to rank pages on the basis or previous relevance or relation of the page with respect to the particular query and user feedback. This paper proposed the work through which crawler give only the relevant or appropriate links by using migrants. The document which gets download are being ranked according to user related field and past knowledge about user visit on a web page and how much time the user spent on it. Whenever we retrieve a webpage the page might be containing images as well some time the images that are fetched are not related to the text associated with it , Image Based Crawling concept is an attempt to get rid of this problem and can affect the Page Relevance score if the image is according to the text or not.

2. PROPOSED WORK:

This work has been done with the provision of satisfying the objective of downloading only the relevant or the matching pages that are according to the searched topic or collection of topics and these pages have the capacity of providing the information related to the user query. In contrast to already existing crawling and ranking techniques Irrelevant or non matching pages are going to be ignored and only the links that consist of large amount of data according to the user search are presented to the user.
Major components of the system are **User Interface, Crawling Manager, Page Relevance Score, Indexer, Migrating agent, Context Manager.**

2.1User Interface

There will exist an interface through which the user will write their queries and ask the web server to serve them with the best possible result. "The user interface is can be defined as the medium where communication between human beings and machines occurs. The goal of this interaction is impressive operation and control of the system i.e. machine on the user's end, and feedback from the machine that is retrieving the information regarding the query, which aids the operator in making operational decisions. It is the part of Web Search Engine establishing communication with the users and allowing them to request and view responses." [4]

2.2 Crawling Manager

It is responsible for providing relevant pages according to prerequisite topic or set of topic it supplied with the set of seed URL's. To earn seed URL the query is submitted to the search engine and from the first n pages the important term appears is stored in D-table. These first n pages are treated as seed URL.

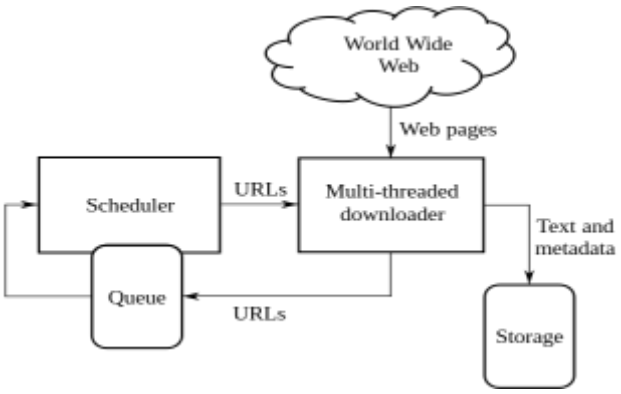


Figure1. **High-level architecture of a standard Web crawler [9]**

2.3 Working of Crawling Manager

It selects the URL from list of seed URL and calls the Migrating Agent [5], along with key table as it contains the keywords which are of high frequency in D Table and this is used to match with web page. The migrant extracts the web page from the web server and then relevancy score of each web page is calculated on the basis of how many terms from Key table appears in web page if value of relevant score of page is on the greater side against the decided threshold score page is considered to be a relevant page and if page is irrelevant migrant return to Crawling Manager. The Crawling manager will also search for the relevant images on the basis of image based crawling.

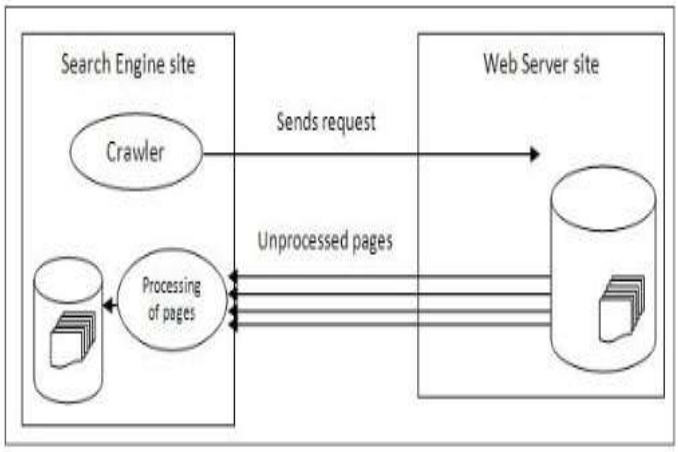
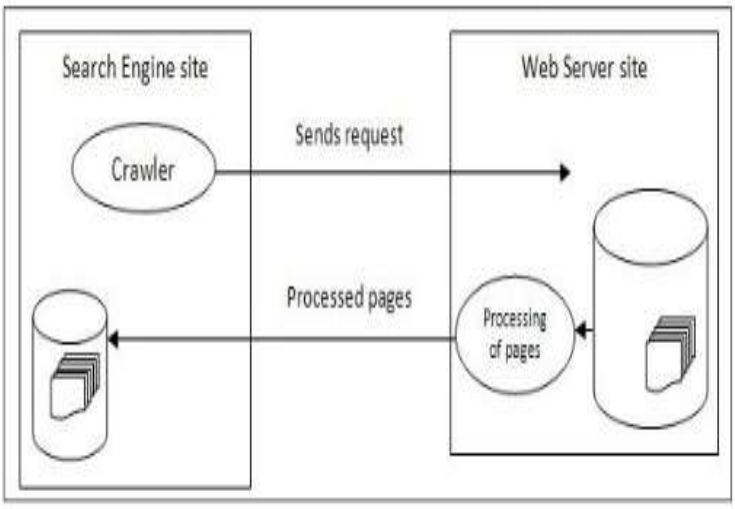


Figure 2.**Traditional Web Crawler [5]**

2.4 Migrating Agents

Migrating agents are computational software processes have the power of visiting large/wide area networks such as the internet, communicating with foreign organization, collecting information on behalf of its boss and reporting back after performing the duties assigned by its master. [5] Migrant returns the relevant web page to Crawler Manager, which further stores in local repository. Repository transferred the link to URL listed called REL URL's and indexed those pages. Using migrants (migrating agents), the procedure of gathering and filtration of web contents can be done at the server side rather than search engine side which can reduce the load on network caused by the traditional Web crawlers. Along with the Migrating Agents there can be another software process that keep on running this process is the process of Context manager that will keep an eye on the context in which the query has been asked, if the pages that are processed are related to context then its fine otherwise a message will be delivered to the Migrating Agent that a particular page is not related to the asked context.

Figure 3. **Web Crawler using Migrating Agents [5]**



Other components are:

2.5 D–Table

There will be existing a structure that will be able to store all the terms that are present in the submitted query, D-Table (short for Data table) contains all the terms that appear in first n-pages for the submitted query. This table will contain all the words that are extracted from the user query after the stop words are removed.
The frequency for each term is also calculated using formula (1)
Fte = total number of presence of te in D-Table / Total number of words present in D-Table …. (1)
Where Fte = frequency of term te ,te=Term

2.6 Combination table

A Combination Table is a data structure that stores synonyms or related words (directly or indirectly) with respect to keywords. Sometimes Combination table can also be called as the Prospective table because it contains all the prospects on which a keyword can be identified.

2.7 Keyword table

It contains the keywords to be verified. This table is sent with the Migrant Agent. Topmost n-terms with highest number of occurrence in D-Table are selected as keywords. Terms from the Combination table that are conventionally or unconventionally associated with each of the keywords selected in D-Table are included in the key-Table.

2.8 Page Indexer

The Page Indexer indexes the web pages in a five table column containing information about the caption, prospective captions for a given caption, URL's where the caption and combination captions appear,
The work of Indexer is to collect the data, analyze the collected data and store the data into the proper format that enables a fast and efficient retrieval of data whenever required."**Stores and indexes information on the retrieved pages**."[1]

2.9 Page Relevance score

Sometimes some WebPages contains all the needed information in the form of pdf, doc files or sometimes in image files as well, so when calculating the page relevance score there might occur a page that might be containing a pdf file or an image file that contains the data related to the user's query. So we will also derive a formula that will match the terms with the keyword-Table by extracting the terms from an Image file, any Pdf file, any Ppt or Word file. The numbers 0.2,0.4,0.1 and 0.3 are taken on the basis of the content importance and also remembering the fact that their summation should be equal to 1.0 , so whatever the number it gets multiplied to should be equal to that number itself.

a)PageRelScore=0.2*TextOfURL+0.4*TextOfMeta +0.1*TextOfHead+0.3*TextOfBody

Another approach that can be applied to find out the usage pattern of the user is to allow the search engine to have a look at the user's search history or the links that have been bookmarked my the user, or the most recent pages that have been visited by the user, or the pages that have been made as the welcome page by the user. The Query that has been entered by the user is taken and the words or the terms are extracted from the query and then the words are matched again the terms that are present there in the Bookmarked or the links present in the history tab and the pages that are most frequently visited by the user gets retrieved or the pages that are related to those pages gets retrieved. Adding User's Personalization and the structure of the page residing on the web can help in better and more prominent Information Retrieval as it is the sole motto of the Crawler and the Search Engine.

b)PageRelScore=0.5*BookmarkLinktext+ 0.3*historylinktext+0.2*newtablinktext

Text of URL: It contains the text of the outgoing links that are associated to a particular page. The text it is containing should be clear and should be relative to the Link it is pointing to.

Text of Meta: Meta Text generally contains all the keywords that are present in the document and these keywords play a very prominent role when the user searches for a topic on the web. It also contains description regarding the document.

Text of Head: The title of the page is placed inside the Head, and this is the area where one of the important keyword related to our document should be presented, it helps the search engine to search the page easily if the title is according to the corresponding document.

Text of Body: This is the area where all the information regarding the document is present or we can say that it is the place where the actual content resides.

Book Mark Link Text: Text associated with the bookmarked links.

History Link Text: Text that is associated with the links present in the History tab.

New Tab Text: The users tend to personalize their search engine home window so the pages that are present there can be taken and from their links the text can be extracted.

Where, TextOfURL= No.of keywords in Keyword-Table that occur in Web Page URL / Total number of terms in TextOfURL

Text Of Meta= No. of keywords in Keyword-Table that occur in Meta tag of web page / Total number of terms in Meta Text

Text Of Head= No. of keywords in Keyword-Table that occur in Head tag of web page / Total number of terms in Text Of Head

Text Of Body= No. of keywords in Keyword-Table that occur in Body tag of web page / Total number of terms in Text Of Body.

Book Mark Link Text: No. of keywords in Keyword-Table that occur in Body tag of web page / Total number of terms in Text Of BookMark Link.

History Link Text: No. of keywords in Keyword-Table that occur in Body tag of web page / Total number of terms in Text Of History tab links.

New Tab Text: No. of keywords in Keyword-Table that occur in Body tag of web page / Total number of terms in Text Of New Tab Links.

In order to define an initial value for the page relevance, let us assume that at least 50% of the contents will get matched to the contents of Keyword Table.

Then by formula:

c)PageRelScore=0.2*TextofURL+ 0.4*TextOfMeta+0.1*TextOfHead + 0.3*TextOfBody[7]
=0.2*(0.5)+0.4*(0.5)+0.1*(0.5)+0.3*(0.5)
=0.1+0.2+0.05+0.15
=0.5
And
d)PageRelScore=0.4*BookMarkLinktext +0.3*HistoryLinktText+0.3*NewTabText
=0.4*0.5+0.3*0.5+0.3*0.5
=0.2+0.15+0.15
=0.5

This value of 0.5+0.5/2= 0.5 mean, of PageRelscore is being used as the initial value and will act as a threshold for all the pages for their relevancy.
If the fetched web page contains any pdf file or ppt file or any document file (word):

e)PageRelScore=0.1*TextofURL + 0.2*TextOfMeta+0.1* TextOfHead +0.3*TextOfBody+0.3(textinpdf+textinppt+textinword)

3. IMAGE BASED WEB CRAWLING:

Search engines are some of the most popular sites on the World-Wide Web. However, most of the search engines today are textual; given one or more key-words they can retrieve Web documents that have those keywords. Since many Web pages have images, selective image search engines for the Web are required. There are two major ways to search for an image. The user can specify an image and the search engine can retrieve images similar to it. The user can also specify keywords and all images relevant to the user specified keywords can be Retrieved [8].Here in this paper I am proposing a new concept for the purpose of Image Based Crawling over the web. In this concept first of all the traditional web crawler will find out the major source of Image from where the images can be taken according to the need and requirement of the user.
1) The rich source of image will be found out by the crawler according to the content that has been specified in the page regarding the resultant images.

2) The second task will take place by performing the Image segmentation using the Image processing techniques. The image that has been retrieved the large number of times and is very much popular among the use in a particular category will be taken and gets segmented.

3) In order to segment an image mostly the Morphological Image Processing will be practiced.
Morphology in image processing is a tool for extracting image components that are useful in the representation and description of region shape, such as boundaries and skeletons. This is middle level of image processing technique in which the input is image but the output is attributes extracted meaning from an image.

4) Once the Boundary of The Image gets extracted the next task is to be performed is to represent the detected shape using the chain codes. The chain codes will create a representation of an image and they help a system (in our case the crawler) that the resulting image is of a particular object and will return only those images that satisfies the shape that is stored in the disk. This will help in retrieving only the relevant images to the query made by the user and won't allow anyone to make a fake label to some other image. For example, the image in the page is of DOG and in the "alt" "title" or "name" attribute of image tag "Tiger" is written.

Above proposed work can be added into the existing crawler or can be used for inventing a new crawler for crawling images only.

4. PROPOSED ALGORITHM:

A step-by step working of the proposed system is given below:

1: First of all a query will be fired by user to any of the popular search engine to retrieve first n pages. The URL's of these pages will serve as Seed URL's.

2: The next step is to remove the Stop words from each page and each term appearing on each page is extracted and stored in D-Table.

Note: The Stop Words are very large in number and affect the page rank of a particular webpage, thus it is the duty of crawler to leave out the stop words while crawling. Examples are of, and, the, etc., that provide no useful information about the documents topic. The process of removing these words is called Stop word removal. Stop-words account for about 15-20% of all words in a typical document. These techniques immensely reduce the size of the search engines index. [4]

3: Then, the Frequency of each term will be calculated using formula (1).

4: The Top n terms having maximum term frequency are selected as keywords. The keywords and their prospective/related terms are stored in key table.

5: Now, The crawler starts with seed URL's. The crawler manager calls the Migrating agents and transfers them to the respective web sites of these
Seed URL's. The migrant also takes the Key table with it for grouping purpose.

6: At the Web-site the Migrant agent retrieves the web page and with the help of HTML Parser parses the document. Then it calculates Page score to determine whether the page is relevant or not. If the retrieved pages contain Images as well then the Images will be checked morphologically and the contents will be matched according to the found images.

7: If the Mean is greater or equal to the initial score that we have already been calculated, the page qualifies to be a relevant or an appropriate page otherwise page is irrelevant page and Migrant Agent ignores the page. The Images that have been checked above if found relevant then they will add positives to the page relevance score otherwise if they are unrelated then the page relevance score will go down hypothetically.

8: The Migrant Agent then transfers the pages that satisfy the user objective and extracted links to a local repository i.e. the database at server side.

9: The Indexer then indexes the pages in the local repository.

10: At the last step the page rank according to the relevancy will be assigned.

5. FUTURE PROSPECTS :

1. The Segmentation of the image can be done on the basis of different-different sizes of the images and from different angles.

2. A Crawler that will only crawl the images over the internet can be developed.

3. Now a day's number of websites are existing over the web, that are having the domain names in Hindi, and when a person types in the name in Hindi the existing crawling technique doesn't really been able to fetch the page and in the URL bar the URL gets written in the form of Unicode's, so this problem can be addressed in near future.[10]

6. ADVANTAGES

1. This proposed technique of using Migrant technology for downloading Relevant document help to reduce load on server.

2. Improve quality of result.

3. Only the relevant pages that are containing the information related to the query are fetched and improve the response time.

4. Image based crawling gets improved as the shape gets checked and reduces the no. of unwanted images gets retrieved , if the name in img tag is mentioned wrong.

7. RESULT

Provide better quality result based on user's preference and provided requirements.

The process of Crawling is implemented In such a way that the pages that are only Relevant will be retrieved.

The Images will be retrieved are more related To the contents that they have been before.

8. CONCLUSION

Web Crawler act as the most important technique of the existing applications that helps in the process of Information Retrieval over the web, thus providing a methodology to improve the quality and working of the Crawler, it can provide much better Results as it are based on user's preference.A technique to improve the retrieval of images is also proposed that alongside with Image processing technique can really prove helpful in the future.

9. REFERENCES

[1] Hema Dubey, Prof. B. N. Roy "An Improved Page Rank Algorithm based on Optimized Normalization Technique" of Department of Computer Science and Engineering Maulana Azad National Institute of Technology Bhopal, India.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank citation ranking: Bring order to the web. Technical report, Stanford University, 1998.

[3] Kleinberg, Jon (December 1999). "Hubs, Authorities, and Communities". Cornell University. Retrieved 2008-11-09.

[4] Ms. Nilima V. Pardakhe[1], Prof. R. R. Keole[2] "An Efficient Approach for Indexing Web PagesUsing Page Ranking Algorithm For The Enhancement Of Web Search Engine Results" of

1) S.G.B.A.U., Amravati, H.V.P.M. College of Engg., Amravati, McMahons Road, Frankston 3199, Australia and

2) Department of Computer Science and Engineering, S.G.B.A.U.,Amravati, H.V.P.M. College of Engg. Amravati,

[5] Managing Volatile Web Contents Using Migrating Agents.

[6] A. Gupta, A. Dixit, A. K. Sharma, "Relevant Document Crawling with Usage Pattern and Domain Profile Based Page Ranking", IEEE, 2013.

[7] ReviewonDocument Crawling With Usage Pattern and Page Ranking Akanksha Upate1 and Surabhi Rathi2

1) Final Year, Computer Science Department, J.D.I.E.T, Yavatmal, India,
2) Final Year, Computer Science Department, J.D.I.E.T, Yavatmal, India,

[8] Crawling for Images on the WWW
Junghoo Cho1 and Sougata Mukherjea
Department of Computer Science, Stanford University.

[9] https://en.wikipedia.org/wiki/Web_crawler#/media/File:WebCrawlerArchitecture.svg

[10] http://hindi.thevoiceofnation.com/technology/hindi-domains-could-soon-become-a-widespread-reality-if-digital-india-club/

International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 56 - 60, 2016, ISSN:- 2319–8656

A Security Model for Virtual Infrastructure in the Cloud

Roya Morshedi
Department of security
Information Engineering,
Central branch, University of
Malek ashtar
Tehran, Iran

Abstract: Accordin to easily manage cloud computing, flexibility and powerful resources on space, provide great potential for improving cost efficiency. Cloud computin capabilities through the efficient use of shared hardware resources increases. Properties mentioned above, incentive agencies and other users of their programs and sevices in this space with a series with a series of threats and risks are also met.
This ensures higher accuracy virtualization and cloud infrastrure components of the virtual machines is. In this regard, particularly for initial design thesis developed a new model called cloud protectionsystem, it is suggested and shown that the proposed model, can increase supply security in the cloud. And packets received by sources and do not be discarded. How to test this architecture, in terms of effectiveness and efficiency in the fight against offensive attacks mentioned above, partly expressed and tools for simulating and measuring the efficiency of the system may be useful, recommended.
Keywords: cloud computing, service levels, virtiuation, model

1. INTRODUCTION

Today, cloud computing is widely used in industry and education. Part of the benefits of a cloud environment, including economic and cost, large capacity, the availability of all places, convenience and access to resources is based on demand, has caused business owners to do their work in this environment instead. References in this environment, on-demand and user requests from multiple sets of resources provided or is released. Preparation of allocation based on demand and are affordable. Consumers, whether ordinary people or organizations no longer need to invest heavily in technology to build their information technology and in this case, users of the resources in the cloud environment and the use to which they pay respectively. On the other hand the cloud could be released as soon as a source by a particular user, use the (reusable resources), resulting in greatly improved utilization of resources. Ease of use is a clear advantage Dygrfzay and other customers to use this space requires special expertise in particular technologies are not cloudy.we can service and Web services, virtualization and multi-tenancy refers reqest via the Internet to customers are clear.the use of physical resources. Virtual multi-processing and process separately from different users on the same physical machines are allocated. However, these resources are logically separate and cause the cloud to be multi-tenant. Despite the benefits that provides a cloud environment, this environment is not free from risk and security risk.[5] Security is one of the biggest barriers that slow the spread of the use of cloud computing.the hold. All processing and data management processes and applications within the field of administrative organization is done. On the other hand the organization of the executive management and infrastructure services and cloud services do not have. security measures cloud service providers, in general, the organizations are transparent and not visible. The presence of a large number of users who do not belong to the organization, increased concerns in the organization. Cloud service providers should rely on users but it may not be mutual trust. The reasons mentioned above, causes the customer to insert their digital assets in the cloud and therefore are doubts about the choice of this medium without relish Grdnd.dr fact, honesty, integrity, confidentiality andIt is clear that despite all the positive and negative aspects, is a comprehensive system and thus increase the protection of nodes in the cloud Karchalsh is controversial.[5] The identification of possible threats and to establish security procedures and services to protect against attacks on the operating system, is very important. Current cloud computing virtualization to provide load balancing between the nodes and physical nodes. That is, if you need to create a new virtual machine or dynamic migrations of virtual machines, load on the network divided between the existing nodes. Virtual machines on the Internet, in the form of different methods can use virtualization technology to filter and separation of data and resources and at the same time, also provide a higher degree of confidence. In particular, virtualization can be used as a security component Grdd.k·h including application virtualization to provide monitoring on a virtual machine, allowing easier management of multiple security cluster and the server mentioned compound.[8] It may seem that the issue is system virtualization for the past decade, however, has a history of more than forty years, basic research has been done in this regard in the 1960s. But in the last decade, significant progress has been made in the virtualization In principle virtualization system using a software layer that surrounds the operating system or its surroundings and the behavior of the input and output The hardware system is expected to provide a Sazd.nrm software to do this is called the hypervisor or virtual machine monitor. It seems that virtual machine monitors are equivalent to the host operating system, but not so in terms of hardware are separated. For virtual systems, virtual environments called virtual machines that may be installed inside or on platforms. Since a virtual machine hardware is not affiliated, may be more virtual machines on the same hardware and the same unit to be installed. A virtual machine is the logical equivalent of a physical hardware and more virtual machines on a single hardware logically separate spaces and can be used in a network as separate systems. The isolated virtual machines mentioned the concept of security is very important. A system-dependent security concepts in the real world is not just a theoretical concept based on factors such as factors and assumptions, implementation details and user preferences can be changed. System virtualization is also not an exception.[6]

2. RELATED WORKS

Issues related to information security and cloud computing in 2011 by "David Frisbee and Vyramvlyab and their colleagues" of engineers, computer engineers and electronics departments Intel was introduced and after the public cloud concepts and challenges in this environment to store storage, transfer and secure computing, a scenario to perform calculations provide securely. Then in 2012, "Shankar Syram and Yvgamankalam" from the University of Tamil Nadu in India that the security problems in the storage and use of open source tools and their resource requests and caching as a solution to the problems mentioned proposed and The use of monitoring tools and virtualization is also useful to know. 2013 can be a turning point in examining issues related to the subject of this paper is that this year at least articles published in journals that went perfectly valid to mention them. In this year, "Gabor pack and Lvnth Attaché and Bnkasa" from the University of Budapest will focus on security issues with hardware virtualization and the concepts relating to the different types of network virtualization and storage services, to expression threats associated with virtualization and Countermeasures notes.explains. Most research in this area in 2015 by "Vasylakv- Ali manifestations and Athanasius" was performed at the University of North Dakota to legal challenges, communication and architecture as well as virtualization refers issues and ultimately "Flavio Lombardi and Roberto Di Pyytrv and colleagues "in 2010, an advanced protection system to maintain the accuracy of the information presented in virtualization.

3.CLOUD COMPUTING

The first question that arises, in relation to the concept of cloud computing and forming part of this environment. In response to this question, the definition provided by the National Institute of Standards and Technology as an academic institution in America, visit We (1 and 5). This definition is widely accepted. It defines cloud computing as follows:

Cloud computing model to provide easy access based on user demand through the network change and configuration set of computing resources (eg, networks, servers, storage, applications and services) that can be accessed with minimal need resource management or the service provider to directly intervene, quickly provided or released (left) is. Essential features are divided into 6 categories that include:

Row property	Row
On-demand access	4.1.1
widespread access network	4.1.2
Resource	4.1.3
The rapid expansion	4.1.4
measured service	4.1.5
Several rental	24.1.6

Table1. The essential characteristics of cloud computing

4. LAYERS AND SERVICES IN CLOUD COMPUTING SERVICE

National Institute of Standards and Technology services provided by cloud services is divided into three categories, namely: 1. software as a service,2. platform as a service 3. Infrastructure as a Service.

4.1 software as a service

This service enables users to use the cloud service provider and run applications on the cloud, it is. Tvanndbh Users access these applications through Web browsers Nmaynd.ayn the possibility of an application does not provide service and only software distribution model to be put on the Internet and users can use the software that do not have ownership of it, according to usage, to pay the costs.

4.2 platform as a service

Applications that is owned by a user, the need for a framework for the implementation and management. This framework includes integrated development environment, the operating system layer resources (time execution engine that will run the software), and is. As the above services by the service provider. This service, control over the user's operating system and applications to the cloud does not pass.

4.3 Infrastructure as a Service

The service provided by the service provider cloud hardware structure implies that includes network, storage space, memory, processors and other computing resources. These resources are available online and via the Internet Bashnd.frahm the cloud service to all service control layer.

5. VIRTUALIZATION

Virtual systems are widely used for a variety of applications to protect physical servers, separate from guest operating systems and debugging software is used. There are many other applications for virtualization and virtualization motives and causes many to choose from there. [6]

An advanced operating system like Windows or Linux is very complex suitable tens of millions of lines of code with the desktop version, and thus a greater level of vulnerability and simply is not possible to prove that they are safe. In addition, the operating system a break point to anything in the system (process and data) to turn and attack the operating system, the entire system was destroyed. It is difficult to secure a spot in the complex point of failure, represent a security risk for data processing in the system. In the result of the permanent reduction of hardware costs, most organizations to achieve different operating conditions based on safety , use of multiple physical systems. [8]

The establishment of a physical system to reduce potential security risks due to a failure point is used, increasing efficiency and flexibility. Each physical device need physical space, cabling, power, cooling and management software is. In addition to the above, due to the physical separation, communication costs such as delays of data transmission and storage should also be added. For some problems, solutions

optimization (such as the storage consecutive to improve access to data and energy management to reduce energy costs), but for some of these problems, the cost is so high that it are not justified. Because of the expenses that each physical systems, one of the most important issues that arise, avoiding systems that are used less. Small system applications within the organization can be seen in two ways: [1] desktop machines that rarely use their full potential (ie, systems that operate at night preservation and maintenance purposes, do not do) 2 - or systems and servers that are not currently active. Many organizations are interested in the efficient use of these systems, and in addition to overhead and low cost, highest efficiency in using them. Virtualization is a method in a hardware system allows multitasking operating system and it is possible to perform multiple operations simultaneously provide and increase the efficiency and reduce the cost.
The benefits of virtualization are:
1.srfh direct and indirect cost savings.
2. Use the optimization of hardware resources and improve efficiency.
3. integration of services in one or more servers, which create a centralized management and high security.
4. accelerate the implementation of the various components and rapid creation of new services in order to increase the organization's business.
5. Support of existing systems and services in the organization.
6. integration of hardware resources.
7. creation and deployment of test environments without disruption and without risk.
8. lower maintenance costs and manpower
9.arayh virtual machines instead of physical machines and run different operating systems on a single physical host.

6.MODEL CLOUD COMPUTING SECURITY

A service provider one or more sample runs in the cloud service that this service can be accessed by a group of final service. For this purpose, providing resources from the cloud service provider's rent. Service users and cloud service provider space are not any physical control over the service level agreements signed with other that specifies how to implement cloud services.
Attacks on cloud systems in the scenario may be divided into two parts:
1. Attacks resources
2. Attacks data
These attacks may include the following:
• Prevent access to resources (denial of service attacks)
• misuse of resources to attack
• steal data or change configuration nodes
• leakage of sensitive information
• attacking the components associated with the structure

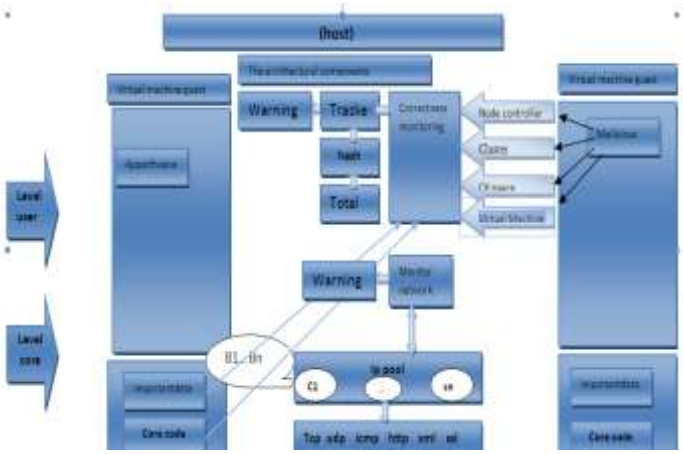
7. Requirements
The main requirements of a system to monitor cloud security we can mention:
Yield property
Efficiency and effectiveness of the system must be able to detect the most common types of attacks and violations of integrity.
System accuracy is better able to avoid negative and positive situations (in this case the attack mistakenly considered to be licensed activities).
Transparency of the system should have minimal visibility into the virtual machine and users are able to detect the presence of possible regulatory system are not
attacker host system, the cloud and other virtual machines should be attempted attack from a guest, be protected and should not be able to change or disable its monitoring system
The ability to expand the system should be expanded on most configuration

Accountability systems should intervene in the cloud and cloud applications, but with data collection and capture must be able to implement policies in response.

8. THE PROPOSED MODEL

The proposed system, to protect the integrity virtual machine guest and distributed computing, as well as to detect and prevent denial of service attacks that use this virtual machine guest using supervision and monitoring infrastructure components and activities in the network. The proposal, developed and protective methods "When I wear him" to protect against intruders and attacks monitored components such as worms and viruses.we will consider.Takes. In fact, guests can target any type of cyber attack and manipulation such as viruses, scripting, and buffer overflows as well. When the image supplied by the user is visiting, security virtual machine is not fully guaranteed and the guest must be possible to track malicious activity, be monitored. In this model, attackers can cloud users or users of their cloud applications. While victims can service providers running in the cloud or cloud infrastructure or other users. One of the common threats and dangers, it is for an attacker to remotely exploit a weakness in a guest system's software. Some attacks exploit cloud services possible. When a malicious person, the legal action of the other cases within the cloud, as mentioned earlier, can be programmed to learn malicious information . Among other possible attacks in the cloud, denial of service attacks, traffic rate is an estimate and should always be monitored traffic on the network.
In order to protect virtual machines and cloud structure of strikes, the key components that can be targeted or they are affected by the monitor (Monitor), we. To enable or disable monitoring on key core or key components and middleware components, capable of detecting any changes might be in the data or the core code. As a result, we can be sure that the integrity of the core and central component of the risk taken place. In addition, in order to monitor entry points to cloud, cloud to the behavior and health components through logging and periodic review of the countervailing executable files and libraries pay. In order to protect from Denial of service attacks, to monitor sent packets on the network and if it was a certain amount of network traffic to prevent service failures due to denial of service attack, to prioritize packets received the query. If the received packet is less important, and the value stored in the buffer delay in executing or discarding it is presented. The next goal, especially when the image cloud provider offers is unreliable, ensure that the program is running forward, is not able to detect external intrusion detection system. However, due to unknown codes in the intrusion detection systems, determine what extent are detected by using a target virtual machine, simply do not accept. In fact, the presence of a monitoring system can be

measured by the performance of certain functions, be detected. The proposed protection system can do the following protection:
1. protected from attacks from outer space is clear.
2. Protect VMs from attacks that it is aligned.
3. protected from attacks that come from the virtual machines.
4. protected from attacks that enter through the network
The proposed system architecture In addition, the architecture "IC Open the" combined. stored. The proposed system has two monitoring systems is at the same time: [1] authenticity observer on key components and data on the network 2. supervisor And can work in two modes: 1. simultaneously V2- for asynchronous notification. honey reqest 2. In the event of an attack when the moment will be warned and prevented from continuing activities. In this system, the coping storage database stored on the host side, which includes the following components: Total coping vital for the components of the architecture of the home and host kernel code and other necessary data and files are used.low priority act. . It all requirements listed in Section 7 meet.



9. PERFORMANCE

As fitness
Hardware and software required
Minimum hardware and software requirements is included in the table shown below:
Details of host 1 host 2
Athlon 64 4400+ Athlon 64 4400+ processor model
Multi 2 2
Memory 4096 4096
Operating system Ubuntu 8.10 (oecp)
Ubuntu 9.10 (eaucal) Ubuntu 8.10 (oecp)
Ubuntu 9.10 (eaucal)
Linux 2.6.30 Linux 2.6.30 kernel
Virtual machine monitor Kvm 88 Kvm 88
Jdvl- hardware and software requirements
The implementation and evaluation of the proposed system requires the following software:
1. virtualization software like kvm or vmware
2. Network simulation software like opnet
3. The network monitoring software like wireshark
4. application traffic generation and simulation attack

10. Problems and challenges

Due to the lack of a code of attacks on the network as well as the relatively low efficiency and high cost of traffic simulators allows simulation of attacks is not complete.
In addition, to obtain the optimum point for different network traffic packets need to experience high performance and accuracy.
Test Mode, once the attack without implementing the proposed system in terms of when and how to diagnose and test again with the implementation of a monitoring system that we repeated attacks. Then there is the possibility of comparisons.

11. WRAP

Cloud computing model to provide easy access based on user demand through the network change and configuration set of computing resources (eg, networks, servers, storage, applications and services) that can be accessed with minimal need resource management or the service provider to directly intervene, quickly provided or released (left) is.Virtual discovery and identification of passive attacks and how to deal with them.

12.REFERENCES

1. Diogo A. B. Fernandes _ Liliana F. B. Soares _ Jo~ao V. Gomes _ M_ario M. Freire _ Pedro R. M. In_acio, Security Issues in Cloud Environments/A Survy, International Journal of Information Security.

2. N.M. Mosharaf Kabir Chowdhury a,1, Raouf Bouta, A survey of network virtualization , Computer Networks.

3. MICHAEL PEARCE, Virtualization: Issues, Security Threats, and Solutions, Acm Computing Surveys.

4. Perez R, van Doorn L, Sailer R. Virtualization and hardware-based security. IEEE Security and Privacy 2008;6(5):24–31.

5. Peter M, Schild H, Lackorzynski A, Warg A. Virtual machines jailed: virtualization in systems with small trusted computing bases. In VDTS ’09: Proceedings of the 1st EuroSys Workshop on virtualization technology for dependable systems, ACM, New York, NY, USA, 2009. p. 18–23.

6. Siebenlist F. Challenges and opportunities for virtualized security in the clouds. In SACMAT ’09: Proceedings of the 14th ACM symposium on access control models and technologies, ACM, New York, NY, USA, 2009. p. 1–2.

7. KELLER, E., SZEFER, J., REXFORD, J., AND LEE, R. B. 2010. Nohype: Virtualized cloud infrastructure without the virtualization. In *Proceedings of the 37th Annual International Symposium on Computer Architecture (ISCA’10)*. 350–361.

8. LI, C., RAGHUNATHAN, A., AND JHA, N. K. 2010. Secure virtual machine execution under an untrusted

management os. In *Proceedings of the IEEE 3rd International Conference on Cloud Computing (CLOUD'10)*. 172–179.

9. STEINBERG, U. AND KAUER, B. 2010. Nova: A microhypervisor-based secure virtualization architecture. In *Proceedings of the 5th European Conference on Computer Systems (EuroSys'10)*. 209–222.

10. SESHADRI, A., LUK, M., QU, N., AND PERRIG, A. 2007. SecVisor: A tiny hypervisor to provide lifetime kernel code integrity for commodity oses. In *Proceedings of the 21st ACM SIGOPS Symposium on Operating Systems Principles.* ACM, 335–350.

11. SHARIF, M. I., LEE, W., CUI, W., AND LANZI, A. 2009. Secure in-vm monitoring using hardware virtualization.In *Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS'09).* ACMPress, 477.

12. SIEBENLIST, F. 2009. Challenges and opportunities for virtualized security in the clouds. In *Proceedings of the 14th ACM Symposium on Access Control Models and Technologies (SACMAT'09).* ACM Press, http://portal.acm.org/citation.cfm?doid=1542207.1542209.

13. WU, X. AND MA, W. 2010. Hypervisor based detection and prevention for packed malware. http://www.ece.tamu.edu/~tristanw/files/Wu Xiaojian Ma Weiqin Report.pdf.

14. YUNIS, M. AND HUGHES, J. 2008. Real security in virtual systems: A proposed model for a comprehensive approach to securing virtualized environments. *Issues Inf. Syst. IX*, 2, 385–395.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 61 - 65, 2016, ISSN:- 2319–8656

Identifying Valid Email Spam Emails Using Decision Tree

Hamoon Takhmiri
Computer Science and Technology
Islamic Azad University
Kish International Branch
Kish Island, Iran

Ali Haroonabadi
Islamic Azad University
Kish International Branch
Kish Island, Iran

Abstract: The increasing use of e-mail and the growing trend of Internet users sending unsolicited bulk e-mail, the need for an anti-spam filtering or have created, Filter large poster have been produced in this area, each with its own method and some parameters are to recognize spam. The advantage of this method is the simultaneous use of two algorithms decision tree ID3 - Mamdani and Naive Bayesian is fuzzy. The first two algorithms are then used to detect spam Bagging approach is to identify spam. In the evaluation of this dataset contains a thousand letters have been analyzed by the software Weka charts provided in spam detection accuracy than previous methods of improvement.

Keywords: Spam; Fuzzy Decision Tree; ID3 Algorithm; Naive Bayesian; Anti-Spam

1. INTRODUCTION

Today, the problem of unintended emails called spam is turned to a serious problem that 80% of these unintended emails refer to spams. Spams make a lot of problems, in other words spams cause the creation of traffic and destroy storage space and authority. Spams cause that users spend a lot of time to divide and clean unintended emails and also cause users' feeling of lack of security. Spams cause some illegal problems such as pornography, pyramidal schemes and economic scams such as phishing sites. In recent years, the increasing popularity and low cost of emails have attracted the attention of direct marketing so that with a promise of winning in lottery and getting valuable prizes, they deceive users. Large lists of email addresses, usually are taken from web pages and archives of news groups, make it possible to send unintended emails to a thousand of receivers without any costs. Users receive large amount of spams that contain anything from holidays to projects of getting wealthy. The term unintended commercial email is used in books too[1]. Spam is used in a wider sense. Spams are annoying for most users because it wastes their time and unsettle their inbox. They also waste users' money by dialing connections, reduce bandwidth and maybe show unimportant subjects with inappropriate contents such as propaganda of vulgar sites. Ferris research institute estimated that economic losses resulting from unintended emails and spams have been over 50 million dollar [2].

2. RELATED WORK

Filters have usually relied on keyword patterns, to be more efficient and prevent the danger of accidental removal of ham messages which are called Ham or allowed messages. These patterns need to be checked with each user's received emails. However, detailed setting of such patterns needs time and proficiency which are unfortunately not always available [3].

Even characteristics of messages will change by the pass of time and need updating of keyword patterns. So, automatic processing of spam messages and allowed messages that have already been received is desirable. Note that text categorization methods can be effective in anti-spam filtering. Unlike most programs of text categorization, indiscriminate mass operation is an unintended message that makes it as spam. The phenomenon can be images, sounds or any other data. The point is that to be able to distinguish between different samples and react based on the type of each sample. Learning usually happens based on one of the following methods: statistically, combination, or neural.

Realizing statistical pattern by assuming that patterns are made based on a random system, is determined based on statistical characteristics of the patterns. Some of the most important reasons of sending spams are economic goals and also advertising for a product, a service or a special idea, deceiving users to use their private information, transmission of a malicious software to the users' computers, creating a temporary failure in email server, making traffic and broadcasting immoral contents [4].

Spams are always changing their contents and forms, so that the anti-spams can't realize them. Some methods to prevent propagation of spams are including:

- economic methods: pay to send emails: like email protocols

legislative methods: such as can-spam law, secure email transfer bed.

- change email transfer protocols and offer alternative protocols such as sending ID.
- control output and input emails
- filtering based on learning (statistics) by using mail features
- detecting a phishing mail (fraud page) by the help of fuzzy classification methods

3. SUGGESTED METHOD

To detect spams better, the first goal is finding behavioral characteristics of the spam, so we need the extraction of data and registration of events of spam's behavior like sender's IP, sending time, amplitude and etc. which are shown in table 1 These data are stored in database, so they are structural data [5].

We can extract the behavioral characteristics of spams from their mail servers. Before the extraction of data, we need the

analysis of characteristics of emails from their reports. Obtaining data technology is chosen to analyze these characteristics, then the main characteristic is obtained and characteristics with less data and weaker connection are deleted. Behavioral features and characteristics of a single email is as follows:

- Customer IP (CIP)
- Receive time (RT)
- Context Length (CL)
- Frequency (FRQ)
- Context Type (CT)
- Protocol Validation (PV)
- Receiver Number (RN)
- Attach number (AN)
- Server IP (SIP)

Table 1 Mail Log Format

Time	IP	Sender	Receiver	Size	Subject	Status
...						
15:18:30	IP1	lzleon79@21cn.com	jsjxy	4987	中京国际物流	spam
15:19:33	IP2	gtfhg65@163.com	gjbjb	890	信息	spam
15:19:35	IP3	cugeuvoxler@euvox.com	0geq00 mthsmejc		Mailbox not exist	spam
15:19:36	IP4	Q0paolino@quadrugby.nl	chk	2442	The wor-ld's large-st selec-tion of online meds available	spam
15:19:39	IP5	30zhangke8@online.ln.cn	chengrw	1303	To chengrw	normal
...						

Features do not exist entirely clear in real world to explain making character for the samples logically and naturally. Data value after preprocessing is as follows:

A) Customer IP (CIP): is used only to calculate the frequency of the transmitter and to extract common pattern of transmitter's behavior, and is not used in calculations of decision tree.

B) Reaching Time (RT): the value of time of day and night is a common value and needs fuzzy making for the degree of transverse (1,0).

C) Context Length (CL): short value, long value and the size of the email are common values and need fuzzy making.

D) Protocol Validation (PV): is Boolean type and when matches with the sender (1) and in case of mismatch (0).

E) Context Type (CT): value in text/Html, multipart. (1) and when tye is text (0).

F) Receiver Number (RN): more value and less value, is a feature of common value and needs fuzzy making.

G) Frequency (FRQ): often or seldom frequency is a feature of common value and needs fuzzy making.

H) Attachment number (AN): more and less value, is a feature of common value and needs fuzzy making. Table 2 lists some examples of after preprocessing results.

Table 2 Attributes From Mail Logs

CIP	RT	CL	FRQ	CT	RN	PV	AN	SIP
...	...	...	...	...	...	...	...	...
IP1	15	4987	2	text	3	valid	0	SIP1
IP2	15	890	4	html	1	valid	1	SIP2
IP3	15	1298	1	html	1	invalid	0	SIP1
IP4	16	2442	3	multipart	2	valid	2	SIP3
...	...	...	...	...	...	...	...	...

Assuming that (A,B) are defined fuzzy subsets in a limited space (F). If A and B are named as a fuzzy rule and recorded as (A→ B) and named as fuzzy condition sets, so B is called fuzzy conclusion sets. The presented knowledge of each fuzzy decision tree shows that the rules are classified as (if - then).

For each path from root to leaves, a rule and a specific path are made. Each value of features is a pair of a part of the piece (and) of a law which is called prior law. The IF part predicts the node of the classification leaf, and so makes the following law (then part). Laws of if-then are for easier understanding, especially when the tree is big[6].

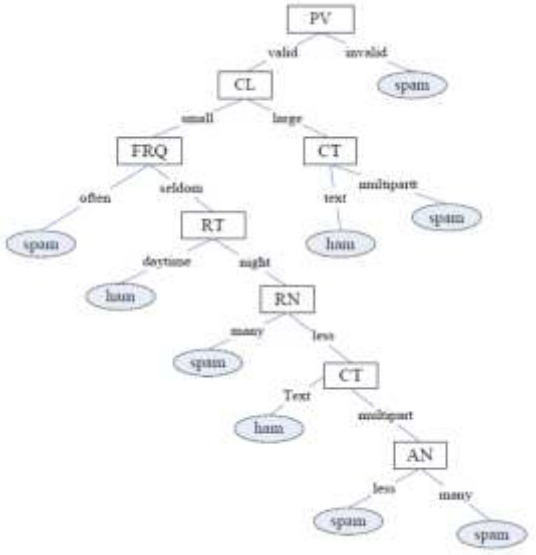


Figure 1 Decision Tree

After examining the decision tree and identifying important features of a mail by the proposed decision tree in figure 1, mamdani's generated decision tree rules are as follows:

1) If the protocol (PV) of email is not reliable, then the email is a spam.

2) If the protocol of email (PV) is valid, context length (CL) is large and context type (CT) is multipart, then the email is a spam.

3) If the protocol of email is valid (PV), context length (CL) is short and frequency (FRQ) is more, then the email is a spam.

4) If the protocol of email is valid (PV), context length (CL) is short, frequency (FRQ) is less or seldom, receive time (RT) is night, and receiver number (RN) is more, then the email is a spam.

5) If the protocol of email is valid (PV), context length (CL) is short, frequency (FRQ) is less or seldom, receive time (RT) is night, receiver number (RN) is less, context type (CT) is multipart, then the email is a spam.

6) If the protocol of email is valid (PV), context length (CL) is short, frequency (FRQ) is less or seldom, receive time (RT) is night, receiver number (RN) is less, context type (CT) is multipart, attachment number (AN) is less or more, then the email is a spam.

7) If the sender's mail server is not valid and reliable, then the email is a spam.

First, spam measures are determined which contain two implicit and tacit parts. Implicit measures are analyzed by Mamdani's fuzzy decision tree, such as protocol type, context length, context type, time, frequency, receiver number, attachment number and etc. Tacit measures are analyzed by Naïve Bayesian method such as frequency of free word repetition, money, three zeros in a row and etc. In fact, the considered data set is a combination of implicit characteristics that are in fuzzy_ Mamdani decision tree and tacit characteristics that are used in Naive Bayesian method. Implicit characteristics of the considered data set are analyzed by decision tree algorithms (ID3) and the results are completed by Fuzzy Mamdani rules [7].

Then tacit characteristics are examined in Naive Bayesian principles and finally, the obtained results from both algorithms are entered in Baking algorithm, that is each mail in a dataset enters the Naive Bayesian and decision tree and in the absence of correct diagnosis (FP and NP) a negative score is registered for the procedure. Finally, the optimal weight may be achieved through trial and error. The bonus rate should also be achieved. This means that the desired class level of the case (or a spam) is divided by the number of spam detection methods. And the result should be divided by the number of mails of the dataset to obtain bonus rate. Mails that are classified correctly are multiplied by bonus rate, and mails that are classified incorrectly are multiplied by bonus rate too. The obtained difference by multiplying the bonus rate in wrong and correct classifying is collected with initial weight (0.5%) This operation is done for Naive Bayesian and decision tree methods and because Naive Bayesian method's threshold is more favorable, it's considered as final threshold. To obtain the ultimate accuracy, each mail is entered in to two Naive Bayesian and decision tree [8].

The output of methods, if both methods have the same results, or in the case of difference, the priority of identification is given to Naive Bayesian method. And to obtain the ultimate accuracy, results are compared with the main class of the mail (spam or ham). When a new mail enters, after the recognition of both methods (Ham=0, spam=1) the output of each method is multiplied by the coefficient obtained from that method, and obtained values are gathered together, for example if just the tree realizes the spam and the other one doesn't realize it, the accuracy is in average and if the response of both methods are the same, for example both detect spam or both do not detect spam, the accuracy is desirable. In the final test by K-Fold method, the data set is divided in to four parts. The first part is for testing and the rest are for learning, in the next step the second part is for testing and the first, third, and forth parts are for learning, then the third part is for testing and the other parts are for learning and after that the forth part is for testing and other parts are for learning [9].

4. RESULT AND DISCUSSION

The dataset that the proposed method is implemented on it contains 1000 emails that 350 (35%) of them are spam and 650 (65%) of them are ham. The last column of this data set is class column and number 1 means spam and 0 means ham. Some examples of keywords for no implicit part of implementation on Naive Bayesian are as follow:

Money, Credit, 000, Internet, Edu, Talent, Free, Make ,# ,$, …

And the other part of this dataset contains implicit characteristics to use for the implementation on fuzzy decision tree, such as:

Sending time, Context type, Context length, Frequency, Receiver number, Sender's number, …

The goal of testing the mentioned dataset is to examine the accuracy of detection of the proposed method and showing a better detection of spams rather than efficiency of Naive Bayesian or decision tree methods. The method is that after analyzing dataset in Naive Bayesian method and extracting levels of efficiency, accuracy and dark bright points and areas, the same data set is analyzed by decision tree and levels of efficiency, accuracy and dark, bright points and areas are extracted, then the obtained results are voted based on Baking method, then the method that has got better comprehension is a priority and its further recognition is collected with the interface of the two methods. To implement in Naive Bayesian method, first the considered data set is implemented in Weka software, then the considered inputs are chosen among fields of dataset, The data set that the proposed method is implemented on it contains 1000 emails that 350 (35%) of them are spam and 650 (65%) of them are ham. The last

column of this dataset is class column and number 1 means spam and 0 means ham. Some examples of keywords for no implicit part of implementation on Naive Bayesian are as follow:

Money, Credit, 000, Internet, Edu, Talent, Free, Make ,# ,$, ...

And the other part of this dataset contains implicit characteristics to use for the implementation on fuzzy decision tree, such as:

Sending time, Context type, Context length, Frequency, Receiver number, Sender's number... ,

The goal of testing the mentioned dataset is to examine the accuracy of detection of the proposed method and showing a better detection of spams rather than efficiency of Naive Bayesian or decision tree methods. The method is that after analyzing dataset in Naive Bayesian method and extracting levels of efficiency, accuracy and dark, bright points and areas, the same data set is analyzed by decision tree and levels of efficiency, accuracy and dark, bright points and areas are extracted, then the obtained results are voted based on Baking method, then the method that has got better comprehension is a priority and its further recognition is collected with the interface of the two methods. To implement in Naive Bayesian method, first the considered data set is implemented in Weka software, then the considered inputs are chosen among fields of dataset [10].

To show the efficiency, the proposed method is discussed with one of these methods. A comparison is done based on accuracy and measurement criteria so that the examined dataset is divided in to ten sections and is examined in groups of 100,200,300,....,1000 mails. The obtained results are compared with the results of spam particle swarm optimization method which contains negative selection method and particle swarm optimization method [11].

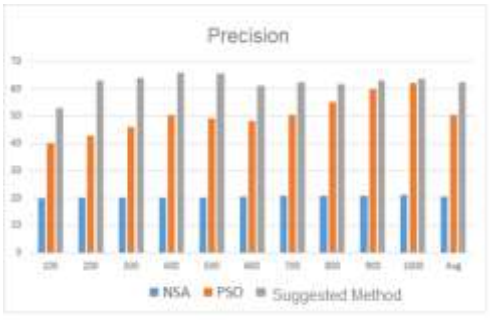


Figure 2 Precision Compare Between Methods

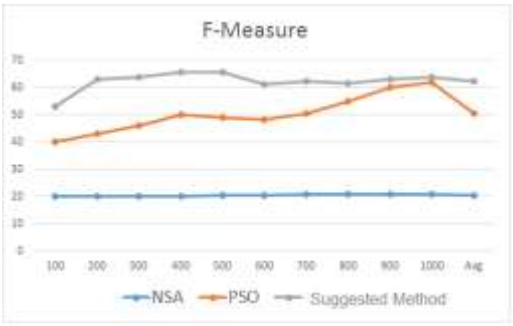


Figure 3 F-Measure

5. CONCLUSION

This method presents a new solution to detect spams by the use of fuzzy decision tree, Naive Bayesian, and Baking voting algorithm to extract spam's behavioral patterns. Because completely clear characteristics don't exist in real world, the degree of crosslinking to explain characters are neutral and rational. Fuzzy decision tree detects spam and ham mails by the use of fuzzy Mamdani rules, then Naive Bayesian method by the use of Bayesian formula does the same operation on chosen dataset, then Baking method by dividing votes in to smaller sections, gaining optimized weight and implementing it on obtained percentages will achieve the level of accuracy and health[12]. The proposed method not only shows a better efficiency in comparison with using each method separately, but also by the use of common interface of spam and ham emails detection (common TPs and TNs of both methods) divides detection in to two categories of reliable and highly reliable. One of the most important items in determining the optimal method of spam detection is minimizing the number of ham mails that are detected as spam mails because finding and deleting a spam among ham mails is easy for the users while finding a ham mail among spam ones is typically difficult and time consuming. To improve accuracy of spam detection results, two methods are used and by the use of Baking voting method and dividing votes, a better spam detection is provided. As mentioned in previous chapter, the comparison of suggested method with some methods that have been done before, shows better performance in terms of obtained accuracy results. Adding a preprocessing fuzzy level to process contents of emails for users by the use of categorizing mails based on content, subject, sender, time, receiver's number, sender's number, and etc. and combining three Naive Bayesian, decision tree, and Baking algorithm methods based on tacit and implicit components of a mail, categorizing has been done based on two methods and voting has been done by Baking algorithm, and false positive and negative rates cause an improvement in the accuracy of statistical filters to detect spams and a decrease in error detection [13].

6. RECOMMENDATIONS AND FUTURE WORK

To improve the proposed method, we can expand branches and leaves of decision tree to enter more details. In fact detailed fuzzy making of a mail includes: sending time,

sending protocol, context length, context type, time zone, number of receivers, frequency, and number of attachments, which increase accuracy performance of decision tree in detecting spams.

Operations such as adding more characteristics to fuzzy Mamdani decision tree and increasing Mamdani's laws improve the efficiency. Adding no implicit details to different parts of a letter such as subject, content, sender, effective keywords in Naive Bayesian method cause the performance improvement of Naive Bayesian method in the field of classifying letters. Finally, the use of both methods in baking algorithm show a better performance percentage. The more the K-Fold divider, the higher the detection accuracy of proposed method is. In other words, the amount of considered K-Fold in proposed algorithm correlates with the accuracy of diagnosis. More attention to details of spam detection and correct classification of mails, results in the increase of accuracy. On the other hand, detection and division of implicit and no implicit characteristics of a mail that each one is detected in its own related method, help a better classification of emails. Note that more attention to details of a mail in detection of a spam will increase accuracy and decrease simplicity and understanding of the method.

7. REFERENCES

[1]. Wu, C.T., Cheng, K.T., Zhu, Q., and Wu, Y.L., 2008, "Using Visual Features For Anti-Spam Filtering", In Proceedings of the IEEE International Conference on Image Processing, Vol. 29, Iss. 1, pp. 63-92.

[2]. Goodman, J., and Rounthwaite, R., 2004, "Stopping Outgoing Spam", In Proceedings of the 5th ACM Conference on Electronic Commerce, pp. 30-39.

[3]. Siponen, M., and Stucke, C., 2006, "Effective Antispam Strategies In Companies: An International Study", In Proceedings of the 39th IEEE Annual Hawaii International Conference on Transaction on Spam Detection, Vol. 6, pp. 245-252.

[4]. Cody, S., Cukier, W., and Nesselroth, E., 2006, "Genres Of Spam: Expectations And Deceptions", In Proceedings of the 39th Annual Hawaii International Conference on System Sciences, Vol. 3, pp. 48-51.

[5]. Golbeck, J., and Hendler, J., 2006, "Reputation Network Analysis For Email Filtering", In Proceedings of the First International Conference on Email and Anti-Spam, pp. 21-23.

[6]. Liang, Z., Jianmin, G., and Jian, H., 2012, "The Research and Design of an Anti-open Junk Mail Relay System", In Proceedings of the First IEEE International Conference on Computer Science and Service System, pp. 1258-1262.

[7]. Feamster, N., and Ramachandran, A., 2006, "Understanding The Network-Level Behavior Of Spammers", In Proceeding of the 3th ACM Conference on Email and Anti-Spam, Vol. 36, Iss. 4, pp. 291-302.

[8]. Lili, D., and Yun, W., 2011, "Research And Design Of ID3 Algorithm Rules-Based Anti-Spam Email Filtering", In Proceedings of the Second IEEE International Conference on Software Engineering and Service Science, pp. 572-575.

[9]. Zhitang, L., and Sheng, Z., 2009, "A Method for Spam Behavior Recognition Based on Fuzzy Decision Tree", In Proceedings of the Ninth IEEE International Conference on Computer and Information Technology , Vol. 2, pp. 236-241.

[10]. Duquenoy, P., Moustakas, E., and Ranganathan, E., 2005, "Combating Spam Through Legislation: A Comparative Analysis Of Us And European Approaches", In Proceedings of the Second International Conference on Email and Anti-Spam,pp. 15-22.

[11]. Jones, L., 2007, "Good Times Virus Hoax FAQ", Available: http://cityscope.net/hoax1.html, [Accesed: Jul. 10, 2015].

[12]. Singhal, A., 2007, "An Overview Of Data Warehouse, Olap And Data Mining Technology", Springer Science Business Media, LLC, Vol. 31, pp. 19-23.

[13]. Ismaila, I., and Selamat, A., 2014, "Improved Email Spam Detection Model With Negative Selection Algorithm And Particle Swarm Optimization", Elsevier Journal of Alliance and Faculty of Computing, Vol. 22, pp. 15-27.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 66 - 70, 2016, ISSN:- 2319–8656

Tech Waste: Environmental Impact and Management

Sangeeta Kumari
School of Information Technology,
MATS University,
Raipur, Chhattisgarh, India-493447

Bincy K Baby
School of Information Technology,
MATS University,
Raipur, Chhattisgarh, India-493447

Abstract: Over the recent years, the global market of electrical and electronic equipment (EEE) has grown rapidly, while the products lifespan has become increasingly shorter. The rapid growth of the electronic and IT industry, current user's culture, increasing rates of usage of techno products have led to disastrous environmental consequences. Most of these technologies are ending up in backlash and recycling centres, posing a new environmental challenge in this 21st century. The presence of hazardous and toxic substances in electronic goods has made tech waste a matter of fear and if not properly managed, it can have unfavourable effects on environment. It has been proven that some of the waste contain many cancer-causing agents. This paper provides a review of the tech waste problems and the need for its appropriate management.

Keywords: Tech Waste; Environmental Impact; Management; Recycling

1. INTRODUCTION

Tech Waste consists of waste generated from used techno devices, electronic appliances which are not fit for their original use and are put-up for recovery, recycling. These wastes consist of various electrical and electronic devices like PCs, mobile phones, including household appliances like refrigerators, air conditioners, television etc. Tech waste contains of thousand different materials many of which are toxic and potentially hazardous to atmosphere and human health [5].

The rapid development of information technologies are being considered as the turning point of human civilization in the afterwards of 20th century and 21st century. The Information Technology has been the strength of the universal economy since 1990s. Frequently increasing production of computer hardware has become major challenges of proper disposal of waste (techno-waste) produced by industries. Recent study focuses on the effect of consumption, dumping and recycling of the electronic waste on the natural atmosphere [8].

During the last decade electronic industry is the largest and fastest growing manufacturing industry. The outcome of its consumer oriented growth combined with fastest product disuse and technological advances are a new environmental objections- the growing danger of tech waste. [3] Information Technology is a developing problem as well as venture opportunity of increasing significance, given the volumes of tech waste being produced and the content of both toxic and valuable substances present in them. In tech waste over 60% is iron, copper, aluminum, gold and other metals, while plastic is about 30% and the hazardous pollutants consist only about 2.70%. [7]

2. CURRENT SCENARIO

2.1 Global Scenario

As per the global e-waste management, Switzerland is the first country to come up with the organized e-waste management system in world. The base of tech waste management system in Switzerland and other developed countries is Extended Producer Responsibility (EPR) and Advance Recycling Fee (ARF). In countries like USA, UK, France & Germany 1.5 to 3 million tons of e-waste are generated annually and are among the largest generators of e-waste. But in these countries also standardized e-waste management processes takes place. Organized and proper e-waste management, from potential sourcing and collection right up to ejection and disposal of material, has ensured that this huge heap of trash turns into a profitable business opportunity. India, China and few African countries have become dumping sites to the developed countries. Due to very drastic environmental standards, the expense of collection, preprocessing, recycling and disposal are very high. There are many countries that started the 'take back' system for their electronic product and have made laws on e-waste management. The US Environment Protection Agency has started National Electronics action Plan in USA to address the various issues related to e-waste. The European Union (EU) has put forward two frameworks for environmental protection from e-waste i.e., WEEE (Waste Electrical and Electronic Equipment) directives and RoHS (Restriction of use of Certain Hazardous Substances) which are also implemented by other countries. As per the EU directives (2003), it is compulsory for all 27 countries of European Union to recycle their waste. [5] In 2014 around 41.8 Mt tech waste generations was there in global quantity. [15] A UN report estimates that 30-50 million ton e-waste is generated yearly worldwide. Nearly 50-80% of e-waste are exported by US for recycling as export is legal in US. The export is due to cheaper labor in developing countries. The recycling and disposal of e-waste in China, India and Pakistan is highly polluted due to the release of toxic chemicals. The lack of responsibility on the part of government for the sustainable disposal of e-wastes have given a way for the development of unorganized sectors for the informal growth of e-waste. [17]

2.2 Indian Scenario

According to Aug 2014 report by Industrial body ASSOCHAM, large usage of gadgets, telecom, IT and appliances is collectively creating nearly 13 lakh tons of e-waste yearly in India. The insight in report is that Delhi-NCR, Mumbai and the IT capital of India, Bangalore collectively produce over 2 lakh tons of e-waste annually. Another January 2015 report from Merchandise and Research has predicted that the Indian e-waste market will grow at the rate of 26.22% CAGR (Compound Annual Growth Rate) during 2014-2019. As per the report, so much technological

waste being generated in the Country, a big portion is handled by the unofficial or unorganized sector using improper processes, which leads to pollution that affect the environment and develop many health hazards [16].

3. ENVIRONMENTAL IMPACT

Roughly each year 40 million metric tons of electronic waste (e-waste) is produced globally, about 13 percent of that weight is recycled in developing countries. According to United Nations Environment Programme (UNEP) 9 million tons of this waste such as discarded televisions, mobile phones, computers, and other electronics are produced by the European Union. And UNEP notes that this estimate is too low. About 50 % to 80% of this e-waste is handled by informal recycling markets in China, Vietnam, and Philippines, often shredding, burning and separating the parts of products in backyards. Emissions from the recycling practices are damaging human health and environment.

Developing countries with fast growing economies handle e-waste from developed countries and their own consumer's. Currently it is estimated that 70 percent of e-waste handled in India is from other nations, but between 2007 and 2020, homely television e-waste will double, computer e-waste will increase 5 times, and cell phones 18 times as per UNEP estimates. The health risks are increased by the recycling practices done by informal sectors. For example, exposure to toxic metals, such as lead, results mainly from burning in open air that is used to retrieve components such as gold.

Burning e-waste release inflammation, it creates fine particulate matter, which cause pulmonary and cardiovascular disease. So several studies in Guiyu, a southeastern City in China, offer insight that the health implications is difficult to isolate due to the informal working conditions, poverty and poor sanitation. Guiyu is the biggest e-waste recycling site in the world, and the city's residents face substantial digestive, neurological, respiratory, and bone problems. For example, 80% children in Guiyu's face respiratory ailments, are at a risk of lead poisoning. E-waste is now the most important global environmental and health issue. Some policy responses have been arisen from European Union, which states source as responsible for e-waste. With this approach, manufacturers are required to eliminate dangerous toxins from production. [11]

Electronic equipment's are made up of a number of components. Some components contains toxic substances which have an adverse impact on the health and environment if not handled properly. These hazards arise due to the improper recycling and disposal processes used. [3]

3.1 Environmental Impact by Different Types of Electronic Components [12]

1. **Cathode ray tubes**
   Cathode ray tube is used in TVs, Computer monitors, ATM, Video cameras, and more. Process used for dismantling and disposal are, breaking and removal of yoke, then dumping. Hazard which affect the atmosphere is such as lead, Barium and other heavy metals extract into the ground water and it releases the toxic phosphor.

2. **Printed circuit board**
   It is a thin plate on which chips and other electronic components are placed. Process used for dismantling and disposal are, computer chips are removed by De-soldering process. Chips are removed then open burning and acid baths to remove final metals. Hazards which effect the atmosphere is air emissions and the discharge of glass, dust tin, lead, beryllium cadmium and mercury into rivers.

3. **Chips and gold plated instrument**
   Process used for dismantling and disposal of this component are, chemical stripping using nitric and hydrochloric acid and burning of chips. Environmental hazards are Hydrocarbons, heavy metals, brominated substances discharged into rivers directly, which is acidifying fish and flora. Tin and lead contamination of surface and groundwater. Hydrocarbons, heavy metals and brominated dioxins are emitted into air.

4. **Plastics**
   Plastics which are used in printers, keyboards, monitors etc. The process used for reuse of the devices are shredding and melting at low temp. Environmental hazards are emission of heavy metals, hydrocarbons and brominated dioxin.

5. **Computer wires**
   Wires which are used in the computer. To reuse the components process used are burning in open air and stripping for copper removal. The environmental hazards are hydrocarbon ashes which released into soil, air and water.

3.2 Effect on Human Being [17]

- The waste element Lead effects the central and peripheral blood system, nervous system, reproduction system and kidney. The source of this waste is Glass panel, Gasket in computer monitors, solder in PCB and other component.
- The waste element Cadmium effects the kidney. The source of this waste is SMD chip registers, semiconductor chips and infra-red detectors.
- The waste element Mercury effects the brain, kidney, and foetus. The source of this waste is electrical and electronic equipment thermostats, relays, sensors, switches, medical equipment, lamps, mobile phone, batteries, flat panel display.
- The waste element Barium causes brain swelling, muscle weakness, damage to heart, liver and spleen. The source of this waste is component used in computer's front panel of a CRT.
- The waste element Beryllium causes lung cancer, skin diseases. The source of this waste is motherboard, finger clips.
- The waste element Toners causes the respiratory treat irritation. The source of this waste is plastic printer cartridge.
- The waste element Hexavalent chromium causes damage to DNA. The source of this waste is untreated steel plant.

4. TECH WASTE MANAGEMENT

4.1 Tech Waste Management Strategies

The best way to handle e-waste is to decrease the volume of e-waste generated. Designers should take care that the product is built for re-use, repair or upgradeability. Stress should be laid on the use of less toxic, easily recoverable and recyclable materials in these technologies such that it can be reused again. Recycling and reuse of materials are the best way of e-waste management. Magnitude of e-waste can be reduced by recovery of metal, plastic, glass and other materials. By doing so energy can be conserved and the environment can be kept free from toxic materials that would have been released otherwise.

Now it is the high time the manufactures, consumers, regulators, state governments, municipal authorities, and policy makers take up the matter sincerely so that the different critical elements addressed in proper manner. It is the time to have an "e-waste policy" and national regulatory framework for the proper management of e-waste. An e-waste policy can be best made by those who understand the in-depth issues caused by e-waste. So it is best for industry to have policy formation with user involvement. By improving the performance of collection and recycling systems the sustainability of e-waste management can be achieved. [7]

4.2 Status of Tech Waste Management in India

Although of a wide range of environment legislation in India there is no specific laws for the guidelines for e-waste or tech waste. As per the Hazardous Waste Rules (1989), unless it is proved to have higher concentration of certain substance e-waste is not treated as hazardous.

Following action /steps has been taken by government to raise the awareness about environmentally sound management of techno waste (CII, 2006):

- Several workshops of e-waste management was organized by the Central Pollution Control Board (CPCB) in association with Toxics Link, CII etc.
- CPCB has initiated the action for rapid estimation of the E-waste generated in major cities of the country.
- A National Working Group has been formed for developing a scheme for E-Waste Management.
- "Environmental Management for Information Technology Industry in India"- A comprehensive technical guide, has been published and circulated widely by the Department of Information Technology (DIT), Ministry of Communication and Information Technology.
- Demonstration projects have also been set up at the Indian Telephone Industries by the DIT for recovery of copper from Printed Circuit Boards.

Despite awareness and readiness for implementing improvements is increasing speedily, the major hurdle to manage the e-waste safely and effectively continues. These include –

- The lack of reliable data that stand as a challenge to policy makers who wish to design an e-waste management strategy and to an industry wishing to make analytical investment decisions.
- Due to absence of a useful take back scheme for consumers only a portion of the e-waste (estimated 10%) finds its way to recyclers.
- The lack of a safe e-waste recycling framework in the formal sector and thus dependence on the capacities of the informal sector pose severe risks to the environmental and human health.
- The existing e-waste recycling systems are purely business-motivated that have come about without any government intercession. Any development in these e-waste sectors will have to be built on the current set-up as the waste collection and pre-processing can be handled efficiently by the informal sector, at the same time offer numerous job opportunities. [9]

4.3 Approaches for E-Waste Management

"Reduce, Reuse, Recycle" should to be adopted for tech waste management. Reduce the generation of e-waste through smart procedure and good maintenance. Reuse still functioning electronic devices by giving or selling it to someone who can still use it. Recycle only those components that cannot be repaired. Use only authorized recyclers for disposing the e-waste products. [10]

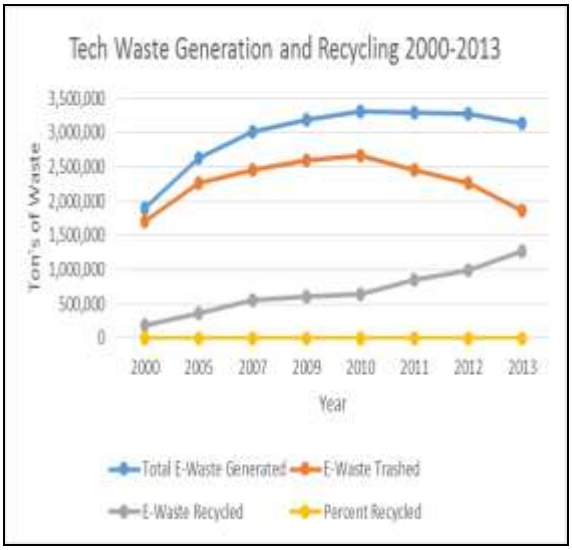


Figure.1 Tech Waste Generation and Recycling 2000-2013 [13]

4.4 Methods for E-Waste Management [16]

1. **Land filling**

   Most commonly used method for disposal of e-waste is land filling. In this method flat surfaces are trenched. Soil is removed from the trenches and waste materials is buried into it, which is covered by thick layer of soil. The degradation process in landfills are very complicated and take a wide time to run. The various hazards of land filling are leaking landfills, leach ate contaminating soil and groundwater, Chemical reactions, vaporization, uncontrolled fires. Thus land filling is not environmentally good treatment for substances, which are volatile and non-biologically degradable (Cd, Hg, CFC), persistent (polychlorinated biphenyls [PCB]) with unknown behavior in a landfill site (brominates flame retardants).

2. **Incineration**

   Incineration is a complete and controlled combustion process, in which specially designed incinerators are

used in which the waste materials are burned at a high temperature (900-1000°C). Advantage of this method is that volume of waste will be reduced and the utilization of the energy content of inflammable materials. Disadvantage is that some escaping substances of flue gas and large amount of residues from gas cleaning and combustion process is emitted into air. Cadmium and mercury are emitted yearly from e-waste incineration plant. Heavy metals are not emitted directly into the atmosphere and it is transferred into slag and exhaust gas residues which can reenter the atmosphere on disposal. So incineration will increase these emissions if no reduction are taken like removal of heavy metals. Incineration Hazards are as follows: Dioxin formation, heavy metal contamination, contamination slag, fly ash, and flue gas, health and safety hazards.

3. Recycling of e-waste

The three types of recycling options for managing plastics from end-of-life of electronics they are Mechanical recycling, chemical recycling and thermal recycling. Recycling process involves dismantling, that is, removal of different parts of electronic devices which contain many dangerous substances like PCB, Hg, separation of plastics, CRT, ferrous segregation and nonferrous metals and printed circuit board. Many precious metals like lead, gold, copper are removed by use of strong acids which affect the atmosphere and human health also. The value of recycling process of any electronic devices is much higher if technologies are used. The persons who recycle the e-waste works in poorly- ventilated enclosed areas without any mask and technical expertise results in exposure to dangerous and slow poisonous chemicals. Devices which can be are recycled are Monitors and CRT, keyboards, laptops, CPUs, chips, mobile phones, compact disks, fax machines, hard drives, floppy disks, telephone boards, modems, connecting wires and cables. Effects due to e-waste recycling are: threat to human health and environment, Lead causes damage to the kidneys, central and peripheral nerve system and blood system in humans, Mercury impacts brain development and functioning.

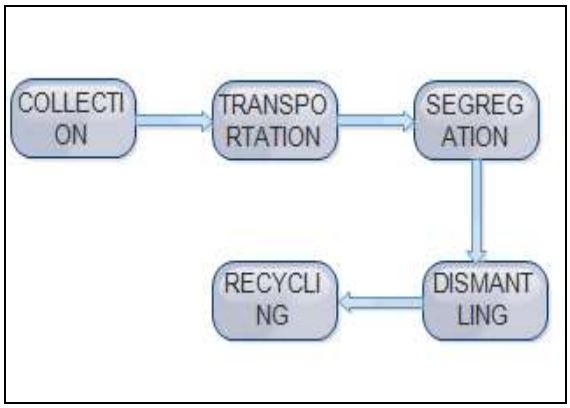


Figure.2 Process of Tech waste Recycling [18]

4. Reuse

It is direct second-hand use or use after slight modifications to the original device. It is used for electronic devices such as computers, cell phones, etc. Inkjet cartridge is also used after refilling. This method also reduces e-waste generation.

5. CONCLUSION

Tech Waste is a growing factor in the world because of the fast development of the electronic and IT industry. The toxic materials present in the techno products is harmful for health and environment. The best way is to have a proper management of the tech waste so that it will turn to profitable product and also a business opportunity for entrepreneur. We should reduce the usage of technology in our day to day life. We can also promote awareness among people about the proper tech waste management. Manufacturers, designers have the responsibility to ensure that the substances used in electronic products can be recycled and reused later such that it does not affect the atmosphere. Improper management of Tech waste is disastrous for human's health, it leads to various types of diseases that are not cured because they are new for doctors. There are some places in India where recycling of Technology products take place. The recycling of disposed techno waste occurs in few places of foreign country like Switzerland, Europe etc.

6. REFERENCES

[1] G. Gaidajis, K. Angelakoglou and D. Aktsoglou, "E-waste: Environmental Problems and Current Management", Journal of Engineering Science and Technology Review 3 (1) (2010) 193-199.

[2] M. Khurrum S. Bhutta, Adnan Omar, and Xiaozhe Yang, "Electronic Waste: A Growing Concern in Today's Environment", Economics Research International Volume 2011 (2011), Article ID 474230,Website:http://www.hindawi.com/journals/ecri/2011/474230/.

[3] Dr. C.Subburaman, "E-Waste Hazardous: Impacts on Environment and Human Health", International Journal of Pharmaceutical & Biological Archives 2012; 3(2):363-367, ISSN 0976-3333.

[4] Ms. Sukeshini Jadhav, "Electronic Waste: A Growing Concern in today's environment sustainability", International Journal of Social Science and Interdisciplinary research, Vol.2 (2), February (2013), ISSN 2277 3630.

[5] Shagun, Ashwani Kush, and Anupam Arora, "Proposed Solution of e-Waste Management", International Journal of Future Computer and Communication, Vol. 2, No. 5, October 2013.

[6] Dr. B. J. Mohite, "Issues and Strategies in Managing E-Waste in India", Indian Journal of Research in Management, Business and Social Sciences (IJRMBSS), I ISSN No. : 2319-6998 I Vol. 1 I Issue 1 I Mar. 2013.

[7] Vijay N.Bhoi and Trupti Shah, "E-Waste: A New Environmental Challenge", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, February 2014, ISSN 2277 128X.

[8] Kurian Joseph, "Electronic Waste Management in India– Issues and Strategies", Proceedings Sardinia 2007, Eleventh International Waste Management and Landfill Symposium S. Margherita di Pula, Cagliari, Italy; 1 - 5 October 2007.

[9] Binay Kumar, "E-Waste – Environment and Human Health Hazards and Management", IRSEE / Prof. (Network Management) / NAIR, Vadodara.

[10] The Human and Environmental Effects of E-Waste, Website: http://www.prb.org/Publications/Articles/2013/e-waste.aspx.
[11] Electronic Waste, Website: https://en.wikipedia.org/wiki/Electronic_waste.
[12] E-Waste in Landfills, Website: http://www.electronicstakeback.com/designed-for-the-dump/e-waste-in-landfills/.
[13] The Global E-Waste Monitor 2014 Quantities, flows and resources, Website: http://i.unu.edu/media/unu.edu/news/52624/UNU-1stGlobal-E-Waste-Monitor-2014-small.pdf.
[14] Nitin Sinha, "E-waste in India: The Current Scenario – 2014 Update", Website: http://attero.in/blogs/e-waste-in-india-the-current-scenario-2014-update/.
[15] Jayapradha Annamalai, "Occupational health hazards related to informal recycling of E-waste in India: An overview", Indian Journal of Occupational; and Environmental Medicine, Year: 2015, Volume: 19, Issue: 1, Page: 61-65.
[16] Dr. Devendra S Verma and Shekhar Agrawal, "E-waste management in India: Problems and Legislations", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 7, July 2014.
[17] E-Waste recycling process in India followed by Exigo, Website: http://exigorecycling.com/e-waste_recycling_process.html.
[18] CII (2006). "E-waste management", Green Business Opportunities, Vol.12, Issue 1, Confederation of Indian Industry, Delhi

International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 71 - 75, 2016, ISSN:- 2319–8656

Social Media its Impact with Positive and Negative Aspects

Shabnoor Siddiqui
Mats University
Raipur (C.G.), India

Tajinder Singh
Mats University
Raipur (C.G.), India

Abstract: Social media is a platform for people to discuss their issues and opinions. Before knowing the aspects of social media people must have to know what is social media? Social media are computer tools that allows people to share or exchange information's, ideas, images, videos and even more with each other through a particular network. In this paper we cover all aspects of social media with its positive and negative effect. Focus is on the particular field like business, education, society and youth. During this paper we describe how these media will affect society in a broad way.

Keywords: social media, business, society, youngsters, education.

1. INTRODUCTION

Now a day's social media has been the important part of one's life from shopping to electronic mails, education and business tool. Social media plays a vital role in transforming people's life style. Social media includes social networking sites and blogs where people can easily connect with each other. Since the emergence of these social networking sites like Twitter and Facebook as key tools for news, journalists and their organizations have performed a high-wire act [1]. These sites have become a day to day routine for the people. Social media has been mainly defined to refer to "the many relatively inexpensive and widely accessible electronic tools that facilitate anyone to publish and access information, collaborate on a common effort, or build relationship" [2].

1. IMPACT OF SOCIAL MEDIA ON VARIOUS FIELDS

1.1 Impact of Social Media on Education

As per the survey of previous research, 90% of college students use social networks. Technology has shown a rapid development by introducing small communication devices and we can use these small communication devices for accessing social networks any time anywhere, as these gadgets include pocket computers, laptops, iPads and even simple mobile phones (which support internet) etc.[5].For the purpose of education social media has been used as an innovative way. Students should be taught to use this tool in a better way, in the educational classes' media just being used for messaging or texting rather than they should learn to figure out how to use these media for good [3]. Social media has increased the quality and rate of collaboration for students. With the help of social media students can easily communicate or share information quickly with each through various social sites like Facebook, Orkut, and Instagram etc. [4]. It is also important for students to do some practical work instead of doing paper work. They can also write blogs for Teachers as well as for themselves to enhance their knowledge skills [3]. Social networking sites also conduct online examination which play an important role to enhance the students' knowledge.

Purpose of Internet Usage	
User	Percentage
Mail	33
Surfing	26.8
Chatting	18.7
Social Networking	17
Other	4.5
Total	100

Fig. usage of social media on education[7]

In the above table 1 it is clear that, internet usage for the respondents was for mailing and surfing the net with 33% and 26% respectively. Mainly two traditional reasons for using Internet i.e. Mailing and Surfing. In India, social networking sites are growing fast to gain popularity but it haven't reached the expectation of global scenario. Just 17% reported social networking sites as their principle reason for Internet usage. Alternating reactions were downloading internet content, purchasing online goods, studying and reading e-books [7]

Membership in social networking sites	
Member of SNS	Percentage
Yes	95.7
No	4.3
Total	100

Fig. Membership in social networking sites for education [7]

Among the Indian youth 95.7% of the members are connected with the social media. These figures are increasing day by day. Whereas only 4.3% of members are not connected with the social media [7].

1.1.1 Positive Effect of Social Media on Education

- Social media gives a way to the students to effectively reach each other in regards to class ventures, bunch assignments or for help on homework assignments [12].
- Many of the students who do not take an interest consistently in class might feel that they can express their thoughts easily on social media [12].
- Teachers may post on social media about class activities, school events, homework assignments which will be very useful to them [12].
- It is seen that social media marketing has been emerging in career option. Social media marketing prepares young workers to become successful marketers.
- The access of social media provides the opportunity for educators to teach good digital citizenship and the use of Internet for productivity [13].

1.1.2 Negative effect of Social Media on Education

- The first concern about the negative effect comes to mind is the kind of distraction to the students present in the class. As teachers were not able to recognize who is paying attention in the classroom [12].
- One of the biggest breakdown of social media in education is the privacy issues like posting personal information on online sites.
- In some of the scenario there were many in appropriate information posted which may lead the students to the wrong side.
- Because of social media students lose their ability to engage themselves for face to face communication.
- Many of the bloggers and writers posts wrong information on social sites which leads the education system to failure.

1.2 Impact of Social Media on Business

Social media is the new buzz area in marketing that includes business, organizations and brands which helps to create news, make friends, make connections and make followers. Business use social media to enhance an organization's performance in various ways such as to accomplish business objectives, increasing annual sales of the organization. Social media provides the benefit as a communication platform that facilitates two way communication between a company and their stock holders [6]. Business can be promoted through various social networking sites. Many of the organization promotes their business by giving advertisement on the social media in order to attract maximum users or customers. Customers can connect and interact with business on a more personal level by using social media. If an organization has established a brand, social media may help this organization to develop the existing brand and give the business a voice. With the help of social media organization can make their strategy to promote their organization.

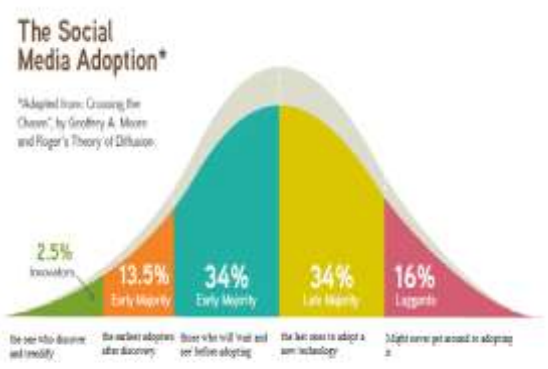


Fig: Social media adaptation [8]

Social media used in various business functions. Some of them are:

Marketing- Marketing is one of the most important and common use of social media in business. It works because today every brand has a target section of online audience.

HR-Is great for identifying and engaging the talent directly.HR helps company to showcase their employee benefits and culture of the company to outside world.

Creative- it share enables art, copy and design teams to invent new ideas which is useful for company to achieve goal.

Operations/strategy- Many of the sites like LinkedIn helps the business by connecting with the experts who can share some strategic plans.

Business Development- Professional networking sites can be used to connect with the clients.

1.2.1 Positive Effect of Social Media on Business

- Social Media helps to better understand their audience by their likes and dislikes [14].
- It helps the business for promotional activities.
- Social networking sites helps to make new customers by providing useful facilities.
- Helps to enhance market insight and stretch out beyond your rivals with online networking [14].
- It also helps to increase awareness among brands and reach with little to no budget [14].

1.2.2 Negative Effect of Social Media on Business

- In business filed social media is not entirely risk free because many of the fans and followers are free to post their opinion on a particular organization, the negative comment can lead the organization to failure.
- Many of the large organization have fallen victim to the hackers.
- The wrong online brand strategy can doom a company, and put at a huge viral social disadvantage[15].
- Getting involved with Social Media is very time consuming. As an organization you should assign a person to always bolster your pages and profile with significant substance [15].
- Most companies have difficulty measuring the results of social media advertising.

1.3 Impact of Social Media on Society

As we all are aware of social media that has an enormous impact on our society[7].Many of the social media sites are most popular on the web. Some social media sites have transformed the way where people communicate and socialize on the web. Social networking sites render the opportunity for people to reconnect with their old friends, colleagues and mates. It also helps people to make new friends, share content, pictures, audios, videos amongst them. Social media also changes the life style of a society.

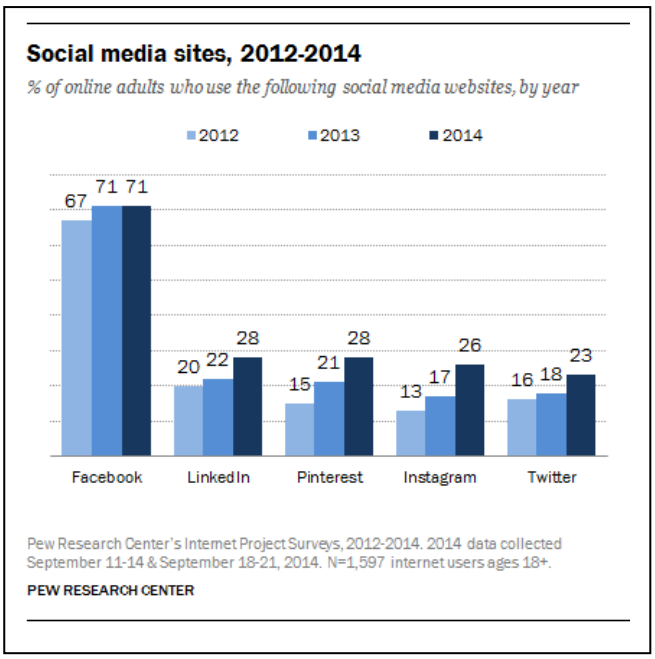


Fig: Usage of social media in the society [9]

According to the survey conducted by the Pew Research Center, in September 2014, 52% of the online adults use two or more social media sites. More than half of the online adults of age 65 and above use 60% of Facebook which represents 31% of all seniors. Half of the internet-using young adult's ages 18-29 use 53% Instagram and half of the Instagram users (49%) use the site daily. The share of internet users with college education using LinkedIn reached 50%. 42% of online women now use the platform, compared with 13% of online men [9].

1.3.1 Positive Effects of Social Media on Society

- Social Media helps to meet people they may not have met outside the social media forums.
- It also helps to share ideas beyond the geographical boundaries.
- It provides open opportunity for all writers and bloggers to connect with their clients.
- Another positive effect of social networking sites is it unite people on a huge platform for the achievement of specific goals. This brings positive change in the society.
- Social media provides awareness among society like campaigns, advertisement articles, promotions which helps the society to be up to date with the current information.

1.3.2 Negative Effects of Social Media on Society

- One of the negative effect of social media is that it make people addicted. People spend lots of time in social networking sites which can divert the concentration and focus from the particular task.
- Social media can easily effect the kids, the reason is sometimes people shares photos, videos on media that contain violence and negative things which can affect the behavior of kids or teenagers.
- It also abuses the society by invading on people's privacy.
- Social lies like family ones also weaken as people spend more time connecting to new people.
- Some people uses their images or videos in social sites that can encourage others to use it false fully.

1.4 Impact of Social Media on Youngsters

Nowadays social media has become a new set of cool tools for involving young peoples. Many young people's day to day life are woven by the social media Youngsters are in conversation and communication with their friends and groups by using different media and devices every day [16]. In past years it was seen that youngsters are in touch with only friends and their groups in schools and colleges. But nowadays youngsters are in contact not only with known friends but also with unknown people through social networking sites, instant messenging etc. [16]. According to BBC news research of 2013 they discuss that 67% Facebook users are very common and well known social media portal consist of the youth and students, so these praise the fact that the youth and student have more focus and relation [11].Throughout the country teenagers frequently use the web, mobile phones, online games to communicate and gather information with each other. As per the survey in California the below table shows that how social media impacts the behavioral health of California's adults [17].

TYPE	EXAMPLE	%TEENS WHO USE SOCIAL MEDIA NATIONALLY
Text Messenging	Cellphone feature	75% of all teens own a cell phone, 88% of cell phone-owning teens text, 72% of all teens use text messaging
Social networking sites	Facebook, MySpace	73% of online teens have used a social networking site
Online video sites	Youtube.com	63% of online teens watch online videos
Online gaming	SecondLife.com	61% of online youth play games online, including multiplayer online games
Blogging with in social networking sites	Facebook or MySpace feature	52% of online teens have commented on a blog

Fig: usage of social media by youth [17]

1.4.1 Positive Effects of Social Media on Youngsters

- Social media helps youngsters to stay connected with each other.
- Useful information can be exchanged over social networking sites.
- Social networking sites can allow teens to find support online that they may lack in traditional relationships, especially for teens [17].
- In a Critical Development period youngsters also go for social networking sites for advice and information.
- Youngsters can look to social media for getting the answers related to their career objectives.

1.4.2 Negative Effects of Social Media on Youngsters

- Today it's not clear that who the "strangers" are especially in the field of social media.
- Kidnapping, murder, robbery can be easily done by sharing details on social media.
- There are many cases registered in police station where adults target young children and lure them into meeting them.
- Mostly youngsters waste lots of time on social sites like chatting which also effects their health.
- Some useless blogs influence youth extremely that they become violent and can take some inappropriate actions.

2. CONCLUSION

As the technology is growing the social media has become the routine for each and every person, peoples are seen addicted with these technology every day. With different fields its impact is different on people. Social media has increased the quality and rate of collaboration for students. Business uses social media to enhance an organization's performance in various ways such as to accomplish business objectives, increasing annual sales of the organization. Youngsters are seen in contact with these media daily .Social media has various merits but it also has some demerits which affect people negatively. False information can lead the education system to failure, in an organization wrong advertisement will affect the productivity, social media can abuse the society by invading on people's privacy, some useless blogs can influence youth that can become violent and can take some inappropriate actions. Use of social media is beneficial but should be used in a limited way without getting addicted.

3. REFERENCES

[1] Aveseh Asough, SOCIAL MEDIA AND ETHICS - The Impact of Social Media on Journalism Ethics, Center for International Media Ethics (CIME),December 2012

[2] https://en.wikipedia.org/wiki.Social_media#References

[3] Gitanjali Kalia Chitkara University, Punjab, A Research Paper on Social media:An Innovative Educational Tool, Issues and Ideas in Education Vol. 1 March 2013 pp. 43–50

[4] www.edudemic.com/social-media-education/

[5] Waqas Tariq, Madiha Mehboob, M. Asfandyar Khan , FaseeUllah, The Impact of Social Media and Social Networks on Education and Students of Pakistan, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012

[6] http://www.business2community.com/social-media/

[7] Dr. M. Neelamalar & Ms. P. Chitra,Dept. of Media Sciences, Anna University Chennai, India, New media and society: A Study on the impact of socialnetworking sites on indian youth, Estudos em Comunicac¸ ˜ao no6, 125-145 Dezembro de 2009

[8] Abhimanyu Shankhdhar, JIMS / Social media and businss /

[9] http://www.pewinternet.org/2015/01/09/social-media-update-2014/pi_2015-01-09_social-media_01/

[10] hubpages.com/technology/effects-of-social-media-on-our-youth

[11] Ghulam Shabir,Yousef Mohammed Yousef Hameed,Ghulam Safdar, Syed Mohammed Farooq Shah Gilani,"the impact of social media on Youth: A case study of Bahawalpur City, Asian Journal of Social Sciences & Humanities Vol. 3(4) November 2014

[12] https://www.schooliseasy.com/2014/02/social-media-in-the-classroom/

[13] http://www.edudemic.com/how-students-benefit-from-using-socialmedia/

[14] http://blog.hootsuite.com/social-media-for-business/

[15] http://www.automatedbuildings.com/news/sep11/columns/110826030404mandrusiak.html

[16] http://www.practicalparticipation.co.uk/yes/what/what_does_it_change

[17] Impact of Social Media on Adolescent Behavioral Health in California,

Source: (Lenhart, 2010) except for Online video sites (Nielsen, 2009) & Online gaming (McAfee, 2010)

International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 76 - 82, 2016, ISSN:- 2319–8656

Excessive Increment in E-Waste System and its Prohibition through Green Computing

Asmita Chawla
School Of Information Technology, Mats University, Raipur(C.G)

Jaswinder Kaur
School Of Information Technology, Mats University, Raipur(C.G)

Abstract— In the current scenario, the information and communication technology have made drastic changes in our daily routine like industries, institution and almost in each field. In today's world there is a large amount of usage of electronic equipments which are giving rise to many problems. The energy consumption from such devices also leading to various global warming issues. At the same time they are leading to many problems like problems of massive amount 0of hazardous waste and other wastes which are generated from electronic equipment

Therefore here we will discuss about various consequences of e-waste , their effects and management of these toxic ad dangerous wastes so as to make the process energy efficient and environment friendly

Keywords— E-waste , Green Computing , Recycling , Dumping, Electronic Waste,

1. INTRODUCTION

In the modern era heavy usage of electronic gadgets during the last two decades has led to increment of a huge amount of e-wastes in soil and environmental pollutants. Thus the major concern nowadays is pollution control and environmental safety. Dumping of electronic wastes has become a major problem in our society. Because these wastes are non biodegradable the gradual reposition of these e-wastes leads to increment of various toxic metals like lead cadmium and pollutes the soil and the ground water. Ground water pollution affects the plants, animal and the humans too as a whole causing severe health problems and disorders.

Therefore, proper management of these electronic wastes has become a crucial demand of the time.

E-Waste

Electronic waste also referred as e-waste describes unwanted electrical or electronic devices. Used electronics which are intended for reuse, resale, salvage, recycling or disposal are also considered as e-waste. Spontaneous processing of electronic waste in all of the countries may cause serious physical and environmental complications, as these countries have limited managerial fault of e-waste processing.

Electronic components such as Cathode Ray Tubes may contain components such as lead, cadmium, beryllium, residents. Even in many countries reprocessing and dumping of e-waste may involve serious risks to the workers and commonality and great care must be taken to avoid uncertain vulnerability in the operations of recycling and leaking of materials such as heavy metals. All other electronic devices & storage media.[1]

Green Computing

Green computing, also termed as green technology, is the environmentally answerable cause of computers and related resources. Such practices include the utilization of energy-efficient central processing units, servers and devices as well as reduced resource consumption and proper dumping of electronic waste (e-waste).[2]

Many IT manufacturers and dealers are continuously spending money in designing energy efficient computer devices, reducing the use of harmful materials and supporting the recyclability of digital devices and paper. Green computing practices came into existence in the year 1992, when the Environmental Protection Agency (EPA) launched the Energy Star program. Green computing aims to attain economic energy and improve the way how computing devices are used. Green IT includes the development of environmentally feasible production practices, energy efficient computers and improved disposal and recycling procedures.[3]

2. CATEGORIES OF E-WASTE

Large Household Appliances	Washing machines, Dryers, Refrigerators, Air-conditioners, etc.
Small Household Appliances	Vacuum cleaners, Coffee Machines, Irons, Toasters, etc
Office, Information & Communication Equipment	PCs, Laptops, Mobiles, Telephones, Fax Machines, Copiers, Printers etc.
Entertainment & Consumer Electronics	Televisions, VCR/DVD/CD players, Hi-Fi sets, Radios, etc
Lighting Equipment	Fluorescent tubes, sodium lamps etc. (Except: Bulbs, Halogen Bulbs)
Electric and Electronic Tools	Drills, Electric saws, Sewing Machines, Lawn Mowers etc. (Except: large stationary tools/machines)
Toys, Leisure, Sports and Recreational Equipment	Electric train sets, coin slot machines, treadmills etc.

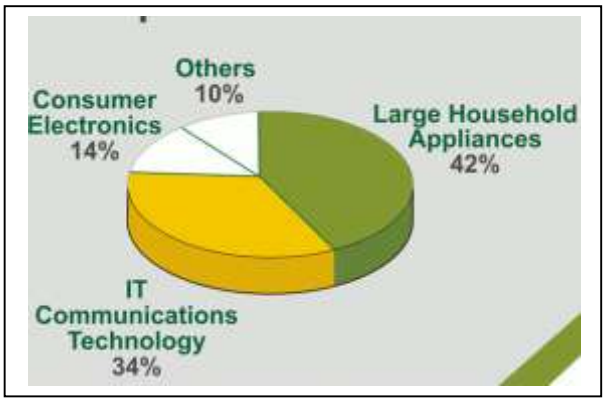


Figure1: Composition of E-Waste[1]

3. IMPACT ON ENVIRONMENT AND HUMAN HEALTH

Electronic wastes can cause comprehensive environmental damage due to the use of harmful materials in the compose of electronic goods. In form or the other

harmful materials such as lead, mercury and hexavalent chromium are present in such wastes which consists of Cathode ray tubes (CRTs), Printed board assemblies, Capacitors, Mercury switches and relays, all kinds of Batteries, Liquid crystal displays abbreviated as LCD, Cartridges of the Photocopy machines, Selenium drums- the photocopier drums and Electrolyte. It is mainly known as, e-waste contains deadly substances such as Lead and Cadmium which are present in circuit boards; lead oxide and Cadmium present in monitor Cathode Ray Tubes (CRTs), Mercury present in switches and flat screen monitors, Cadmium exists in computer batteries; polychlorinated biphenyls (PCBs) present in capacitors and transformers which are very older, and brominated flame choke off on printed circuit boards, plastic casing, cables and Polyvinyl chloride abbreviated as PVC, cable insulation that discharge highly toxic dioxins and furans when burned to regain Copper from the wires.

All electronic components contain printed circuit boards which are very hazardous because they contain lead, brominated flame retardants (which is typically 5-10 % by weight) and antimony oxide, which is also available as a flame retardant (which is typically 1-2% by weight). Land filling of e wastes can be pointed towards extracting of lead into the ground water. If the CRT is crumbled and burned, it emits harmful fumes into the air. These products contain several rechargeable batteries, all of which contain toxic substances that can pollute the environment when burnt in incinerators or disposed off in landfills.

The cadmium from one cell phone battery pollutes 600 m3 of water. The amount of cadmium in landfill sites is authoritative, and considerable lethal contamination is caused by the unavoidable effects of cadmium leaking into the surrounding soil. Because plastics are highly combustible, the wiring board and electronic products contain brominated flame residents, which are clearly damaging to the health of humans and many living organisms as well as surroundings too.

Health Risks

Recycling of havoc carries health risks if relevant caution is not taken. Workers those who are working with waste which contains chemical and metals may sense a sensitivity to hazardous substances and have major health problems at the range of physical disorderliness, inabilities etc. Toxic exposure even sometimes may become poisonous. Therefore, dumping of healthcare wastes and toxic metal wastes require special attention in order to abstain major health hazards.

4. PROBLEMS ARISING THROUGH E-WASTE

With the rapid-advancement in today's society, electronic appliances of all forms have rapidly unify themselves as a necessity in our daily lives. The TV sets that entertain us, to the GPRS that navigate us; from the headphones in our ears, to the, from the cell phones we communicate through, to the computers we work on. Eagerly, we scramble to proudly hold the latest and greatest.[5]

The amount of electronic devices junked globally has increased recently, with 20-50 million tones generated each year.

Electronic waste (e-waste) at present makes up five percent of all municipal solid waste worldwide, nearly the same amount as all plastic packaging does, but it is much more poisonous than that. Not only developed countries produce e-waste, Asia shelve an estimated 12 million tonnes of e-waste every year.

Figure: Showing E-Waste in India

5. SOURCES OF E-WASTE

Home	PC Television Radio Cell Phones Microwave Oven Washing Machine Electronic Iron
Hospitals	PC Monitors ECG Devices Microscope Incubator
Government	PC CPU FAX Machine Scanner Tube Lights

	Air Conditions
Private Sectors **(Restaurants, Industries)**	PC Boiler Mixer Signal Generator Incubator

6. E-WASTE IN INDIA

As there is no independent collection of e-waste in India, there is no crystal clear data on the quantity developed and dumped on each year and the resulting amount of environmental exposure. The preferred method to get rid of extinct electronic items in India is to get them in interchange from vendors. The business division is estimated to report for 78% of all installed computers in India. Discarded computers from the business zone are sell-off. Sometimes educational institutes or charitable institutions collect old computers for their reuse. It is estimated that the total number of discarded personal computers emerge each year from business and individual households in India will be around 1.38 million. According to a survey of fraternity of Indian Industries, the total waste generated by discarded or damaged electronic and electrical devices in India has been estimated to be 1,46,000 tons per year. The results of a field report conducted in the Chennai, a city of India to determine the average usage and life of the personal computers (PCs), television (TV) and cell phones showed that the average homely management of the PC ranges from 0.39 to 1.70 build upon the income class. Television sets ranges from 1.07 to 1.78 and for mobile phones it ranges from 0.88 to 1.70. The low-income households use the Computers for 5.94 years, Televisions for 8.16 years and the cell phones for 2.34 years while, the upper income class people uses the Computers for 3.21 years, Television for 5.13 years and cell phones for 1.63 years. Although the per-head waste production in India is still comparatively small, the complete volume of wastes generated will be very high. Further, it is increasing at a faster rate. The growth rate of the mobile phones (80%) are very high as compared to that of PC (20%) and TV (18%) as people use mobile phones much more than PCs and TVs. The public alertness on e-wastes and the eagerness of public to pay for e-waste management as computed during the study based on an organized census revealed that about 50% of the public are informed about the environmental and health impacts of the electronic devices. The eagerness of public to pay for e-waste management varies from 3.57% to 5.92% of the product cost for Computers, 3.94 % to 5.95 % for TV and 3.4 % to 5 % for the cell phones.

Additionally appreciable quantities of e-waste are reported to be intended. However, no adequate figures available on how generous are these e-waste streams. The government trade data does not categorize between new and old computers and external parts and so it is very complicated to track what measure of imports is used electronic goods.[4]

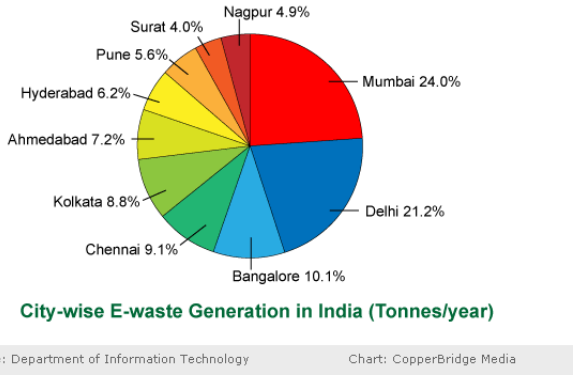


Figure: City wise E-Waste Generation in India[12]

7. ELECTRONICS ARE DIFFICULT TO RECYCLE

Recycling electronics is not as easy as recycling papers. These products are not simple to recycle. Proper and safe recycling often are very expensive than the materials.

Electronics which are not designed for recycling process Materials used and physical designs make recycling difficult. While companies are putting affirmation to inquire "green electronics," we are miles away from the green products which should be used.

Electronics which contain many toxic/harmful materials Screens of Computers and Television sets made with tubes (not the flat panels one) which contain very much amount of lead in them. Most of the flat panel monitors and Television which are being recycled now contain less lead as compared to earlier, but more mercury, from their respected mercury lamps. About 40% of the heavy metals, including lead, mercury as well as cadmium, in landfills rise from electronic equipment junks.

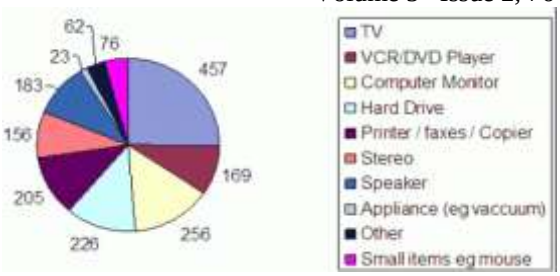


Figure: E-Waste Recycled since 1/8/09 (number of items)[12]

8. FORTHCOMING TRENDS

The global e-waste production is probably to raise due to the economic development and the available technologies since the increased GDP points to increased buying of electronics and basically increased e-waste manufacture.
The increasing economic growth is expected to replicate higher e-waste production. On the contrary, it is expected that specific changes in the technology and the utilization habits are likely to decrease the worldwide e-waste production, since customers may support more convenient PC solutions having 1-3 kg standard weight compared to the immobile computer weighing 25 kg, or the immobile computers is projected to be prepared with LCD (Liquid Crystal Display) screens instead of the older CRTs (Cathode Ray Tube).[8]

9. E-WASTE AS INFORMATION SECURITY

The safety of the information on your electronic devices are the most beneficial issue when choosing an electronics recycler E-waste presents a possible security risk to individuals and exporting countries. Hard drives that are not correctly erased before the PC is disposed of can be reopened, revealing sensitive information. Credit card numbers, confidential financial data, account details, and records of online dealings can be accessed by most enthusiastic individuals. Ordered criminals usually look up for the drives for information to use in confined scams.

10. GREEN COMPUTING AN ECO FRIENDLY APPROACH TOWARDS E-WASTE

Green computing is a fresh technology that is now in attention of business, industries for the energy performance and to dispose E-waste in an effective and risk-free way. They now came to understand that going green is in best interest, both in terms of public affairs and cheap costs. Force of computing was initially on quicker analysis and speedier calculation and solving of more difficult problems. But in the latest past Green computing has got vast importance and that is reaching of energy efficiency, minimization of power utilization of e-equipment.

It has also given highest awareness to minimization of e-waste and use of non-hazardous materials in preparation of resourceful computers and Electronics. The main goal of this technology is to study and apply computing resources efficiently and naturally. Maximizing the energy efficiency and to support biodegradability which are the prime focus of this technology. Due to pollutants generated by it and the regular increment in rates, energy utilization is causing a serious environmental and monetary problems. On the subject of energy efficiency, a branch of Green IT named energy-aware computing has emerged. Green computing is very much necessary for the forthcoming world. It is required to make our self as well as our environment healthy and fit. It can be defined as sensibly utilizing the resources presented. Many computers are formed from many poisonous substances like cadmium, mercury and other harmful objects. While disposing off the computers, it will lead to pollution and affect the environment to a greater extent. This field encircles a broad range from new generation techniques to the study of higher materials to be used in our daily life. Bringing it to practice will deal with many problems that are being a risk to human life and our environment too. The impact of the poisonous wastes that are produced by us from throwing our old computers and peripherals which leads to land pollution. The computers have the power hogs that produce pollution by the energy they absorb for their processes.[9]

GREEN COMPUTING SOLUTIONS AS OUR MAJOR GOAL

Developing sustainable green-computing plans

This involves active contribution of all the citizens those who are linked with the organisations from the finest levels to the ground level. Organizational policies and catalogues needs to be organized, containing obligatory guidelines, government policies, ―green-recommendation, list of eco-friendly and non-recyclable items. The greatest practices and procedures which should aim at decline of usage of non-conventional resources, by falling usage of paper and recycling of old devices and systems in order to abolish e-wastes from the organizations.

Recycle and Reuse

Rejected, used or unwanted electronic tools in a convenient and environmentally answerable manner. Computers have contaminant metals and pollutants that can release dangerous emissions in the environment. Computers should never, ever be eliminated in landfills. Recycle them instead through producer programs such as recycling facilities in your society. Or give away still-working computers to a non-profitable organization.

Purchase products which are environmentally green

Buy products which are labeled as green and safe for you as well as the environment. These products help to reduce the deprivation caused by the energy-consumption to the environment. For these consumers should be motivated to buy products which are environmentally sound. Clear and unblemished criteria must be set for the desire of green-products. Creators should be also involved and given appropriate credits for the process of manufacturing products which are beneficial for the environment. Buy the Electronic Product Environmental Assessment Tool registered products. EPEAT is a acquisition tool promoted by the nonprofit Green Electronics Council.

Minimizing consumption of paper

There are many ways to nullify the usage of paper. With computers being more popular than any other thing today all jobs can be done on the accumulator. Complicated modes of communications like e-mail, free-messaging, and other social sites that have brought communication to your approach. Moreover paper is being saved by the industries, as many industries are trying to change themselves to ―paper-less-mode day by day.

Conservation of energy

All electronic devices show the consumption of energy which has been taken from non-renewable energy resources. So adopt the suitable strategy and techniques so as to preserve energy so that it can be recycled when there is the actual need.

11. FORTHCOMING TRENDS

The forthcoming trend of Green Computing is going to be based on competence, rather than decrementation in consumption .

The primary focus of Green IT is in the organization's self interest in energy cost decrement, at Data Centers and at desktops too, and the conclusion of which is parallel Reduction in carbon generation. The secondary focus of Green IT needs to be focus ahead of the energy use in the Data Center and the focus should be on modernization and improving arrangement with overall corporate social responsibility efforts. The secondary focus will insist the development of Green Computing strategies. The idea of feasibility addresses the topic of business value formation while ensuring that long-term environmental resources are not affected. There are few efforts, which all companies should take care of

A. Certifications for Green Products

B. Cloud Computing

C. Product Longevity

D. Power Management tools

E. Leveraging Unused Computer Resource

F. Data Compression

G. Application

12. ADVANTAGES AND DISADVANTAGES OF GREEN COMPUTING

Advantages:

- Energy saving
- Environmentally Friendly
- Cost-effective (pays over time
- Save more money per year
- can give you a tax right off

Disadvantages:

- High expenditures
- Not readily available
- Still in experimental stages
- Sacrifice performance for battery life
- Not for everyone

13. CONCLUSION

"Technology is not a passive observer, but it is an active subscriber in obtaining the goals of Green Computing."

As Consumers we have only taken care of speed, price and performance factors of the electronic appliances and gadgets but the thing which we haven't cared about at all is their ecological impacts. But with the mushrooming concern on environment and surroundings, people have started thinking about safer and greener concepts.

Division of E-waste into specific branches at collection stage is clearly an helpful approach for providing consequent valuable recycling and reuse.

At present various companies have developed many technologies through it can recycle wastes and does not use any chemicals along with it. Efforts are being pushed up by IT sector to achieve green computing by reduction and recycling of resources. The rules and regulations of government are driving dealers to behave green, do green, go green, think green and act green to adopt green computing.

All these efforts are still in process mainly to reduce the E-waste but the future of Green Computing will be depending on efficiency and the green products.

14. REFERENCES

[1] https://en.wikipedia.org/wiki/Electronic_waste

[2] http://searchdatacenter.techtarget.com/definition/green-computing

[3] [https://www.techopedia.com/definition/14753/green-computing

[4] [ELECTRONIC WASTE MANAGEMENT IN INDIA–ISSUES AND STRATEGIES KURIAN JOSEPH Centre for Environmental Studies, Anna University, Chennai, India

[5] http://www.imselectronics.com/e-waste-problem/

[6] .http://www.greenpeace.org/international/en/campaigns/detox/electronics/the-e-waste- problem/

[7] .https://en.wikipedia.org/wiki/Electronic_waste

[8] **E**-waste: Environmental Problems and Current Management,G. Gaidajis*, K. Angelakoglou and D. Aktsoglou**,** Journal of Engineering Science and Technology Review 3 (1) (2010) 193-199

[9] Green Computing: Issues on the Monitor of Personal Computers, A.Mala,, C.UmaRani, L.Ganesan, Research Inventy: International Journal Of Engineering And Science Vol.3, Issue 2 (May 2013), PP 31-36 Issn(e): 2278-4721, Issn(p):2319-6483

[10] Green Computing: From Current to Future Trends, Tariq Rahim Soomro and Muhammad Sarwar, World Academy of Science Engineering and TechnologyInternational Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering Vol:6, No:3, 2012

[11]https://www.google.co.in/search?q=e+waste+management+images&tbm=isch&imgil=OJNofjDnNatUoM%253A%253B_EFiDtSK8a-6aM%253Bhttp%25253A%25252F%25252Fe-wastage.weebly.com%25252F&source=iu&pf=m&fir=OJNofjDnNatUoM%253A%252C_EFiDtSK8a-6aM%252C_&usg=__-McDZnTTo3NPU5pk-fZOq0hv3QY%3D&biw=1517&bih=741&dpr=0.9&ved=0ahUKEwj9zvi2uL3KAhWNWY4KHVLUCpYQyjcILw&ei=SCOiVv2KE42zuQTSqKuwCQ#imgrc=OJNofjDnNatUoM%3A

[12]https://www.google.co.in/search?q=e+waste+management+images&tbm=isch&imgil=OJNofjDnNatUoM%253A%253B_EFiDtSK8a-6aM%253Bhttp%25253A%25252F%25252Fe-wastage.weebly.com%25252F&source=iu&pf=m&fir=OJNofjDnNatUoM%253A%252C_EFiDtSK8a-6aM%252C_&usg=__-McDZnTTo3NPU5pk-fZOq0hv3QY%3D&biw=1517&bih=741&dpr=0.9&ved=0ahUKEwj9zvi2uL3KAhWNWY4KHVLUCpYQyjcILw&ei=SCOiVv2KE42zuQTSqKuwCQ#imgrc=7cqcENnMP2OWdM%3

International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 83 - 87, 2016, ISSN:- 2319–8656

Exploration and Supremacy of Li-Fi over Wi-Fi

Jessemine Antony
School Of Information Technology,
MATS University
Raipur, India

Prakash Verma
School Of Information Technology,
MATS University
Raipur, India

Abstract: To accomplish the work, the need of internet either through wired or wireless network is increasing nowadays. While using wireless network i.e. Wi-Fi, many issues are arising related to speed due to which the speed of transmitting data goes relatively slow as many devices gets connected. To remedy this, Harald Hass invented technology named Li-Fi which he terms as- Data through Illumination, where the data is transferred through an LED bulb which is 1000 times faster than Wi-Fi. This technology has now become the part of VLC as this technology is performed by using white LED light bulbs.

Keywords: Li-Fi (Light Fidelity), LED (Light Emitting Diode), Wi-Fi (Wireless Fidelity), VLC (Visible Light Communication).

1. INTRODUCTION

One of the most important activities in the current time of this fast moving world is transfer of data and information. As we need faster transfer of data, we can't avoid Wi-Fi in today's world. Though Wi-Fi is providing benefit by giving access to many devices and transferring the data at a high speed, it also turns up as a drawback of it. As the various types of devices such as Ipads, computers and many more are frequently increasing day by day, the limited bandwidth leads to the reduction in the speed of the data transfer. So to overcome this problem Li-Fi technology was introduced by **Harold Hass** to transfer the data at a high speed using VLC(Visible Light Communication). Well to simplify it more clearly, this technology can be thought of as a light based Wi-Fi. No one would have ever imagined that this invention will be used not only to illuminate the houses but also to transmit the data at a high speed.

Figure 1. Li-Fi Bulb [1]

2. Li-FI DESIGN

The architecture of Li-Fi consists of many LED lights, bulbs and lamps with many wireless devices such as cell phones, idea-pad and other devices supporting internet on it. Important constituents that need to be considered while the designing of Li-Fi are as follows: [2]

a) Presence of light
b) Line of Sight (LoS)
c) Fluorescent light and LED can be used for better performance
d) Photo detector

As shown in the below figure, the flow of content must be properly incorporated with server and internet network, so that it is easily possible to work effectively.

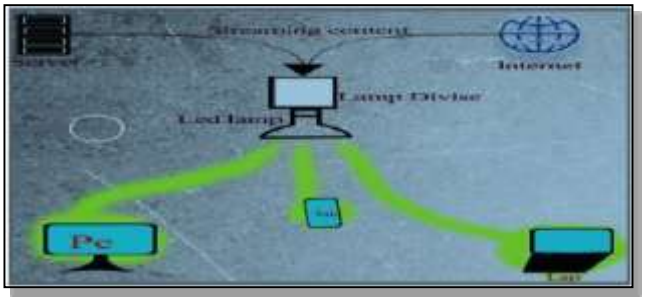


Figure 2. Architecture of Li-Fi[2]

3. WORKING OF Li-Fi

Before describing about the working of Li-Fi, we need to know the requirement of Li-Fi. Dynamically as the lifestyle is developing in respect to the time, the need of using the internet is also increasing. As the usage is incrementing day-by-day, the performance of Wi-Fi is degrading as many types of devices gets connected at a time which reduces the speed and power of it. To surmount this problem, Li-Fi technology was introduced which is basically implemented using LED lights. It renders logic that if LED is on then it transmits a digital signal 1 and if LED is off then it transmits a digital signal 0. The large bright LED lights can be switched off and on very rapidly or say quickly which gives nice opportunities for transmission of data through light [3].

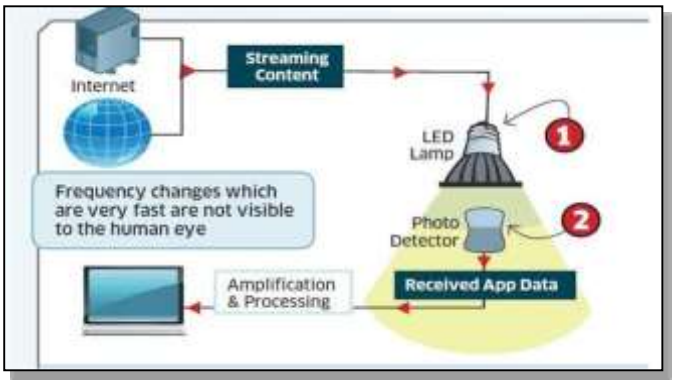


Figure 3. Working of Li-Fi [4]

Therefore all that is required is some LED lights and a controller that code the data into those LEDs [2]. The information encoding is possible in lights by monitoring and identifying the rate variance which all depends on the flickering of LEDs i.e. on and off to code the data accordingly which will pass binary strings i.e. 0s and 1s [5].On one corner there will be a light emitter i.e. the LED and a photo detector i.e. a light sensor that converts light into current on the other end. The photo detector senses the light and converts the light into current by registering 1 when the LED is on and 0 when the LED is off. The intensity of LED is so high which when regulated gets impossible for a human eye to detect. As it is undetectable by a human eye it seems to be constant. These light waves gets undetectable as the light emitting diodes can be switched off and on very quickly which causes the light source to appear in a continuous state even though being in the flickering state. VLC (Visible Light Communication) is the method used for transmitting the information by using rapid pulses of light [6].

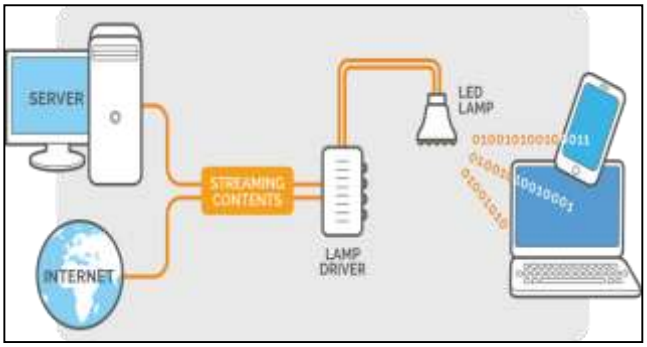


Figure 4. Working of Li-Fi [7]

- ***VLC (Visible Light Communication)***

VLC is the wireless technology for next generation that uses light emitting diodes (LEDs) that offers multiple roles of illumination as well as data transmisison.Usage of fast pulses of light is there which helps in transmitting information wirelessly. Data like audio, video and other types of data can be transmitted at a high speed using LED lights. It can be said as a data communication medium that uses visible light between 400 THz (780 nm) and 800 THz (375 nm) that act as an optical carrier for data transmission and illumination. The use of fast pulses of light helps in transmitting information wirelessly. [8], [9]

The major constituents of VLC are firstly the LEDs that is an essential element to be used and secondly the photodiode i.e. the photo detector that senses light and converts the light into current which helps in transferring the information with more accuracy. To define it more clearly, by regulating the data signals in the form of lights, the LED bulbs can be referred to as a communication source. The bandwidth is comparatively 1000 times more in size which makes it easy in transmission of data with different data channels at high speed. Adding up an advantage that these visible lights are not harmful to vision which is therefore a necessary part of the infrastructure and is easily available and accessible[9].

Therefore LED lights are preferred in working with Visible Light Communication (VLC).

Figure 5. Photo Detector (sensor) [10]

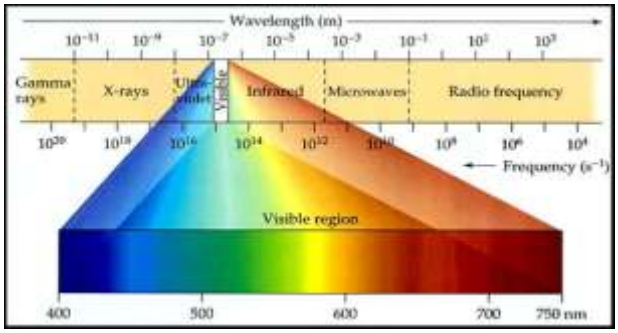


Figure 5. Visible Light Spectrum [14]

to it whereas light waves with higher frequency and shorter.

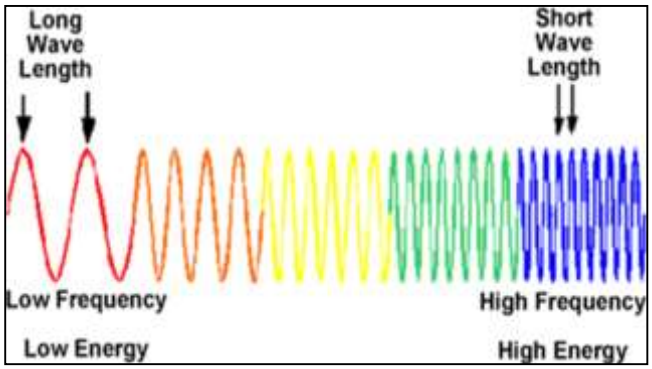


Figure 6: Relation between Wavelength and Frequency [15]

4. ADVANTAGES OF Li-Fi

Li-Fi technology is basically based upon LED lights for the transmission of data through illumination effectively and efficiently. Any kind of information like movies, games, images and many more can be downloaded in a very less time. By providing benefits for this technology, the justification of the superiority of Li-Fi over Wi-Fi is defined below:[11]

a. CAPACITY

The bandwidth of light is 1000 times wider than the radio waves bandwidth which enables the transfer of data effectively.

b. EFFICIENCY

Efficiency in terms of light refers to the minimum utilization of energy consumed by LED lights which is also cheaper and efficient.

c. AVAILABILITY

Presence of light means Li-Fi is available but for more efficiency in this technology if LED bulbs will be set then there will be proper transmission of data.

d. SECURITY

Unlike Wi-Fi, light waves cannot get across through walls so no worries of getting misused.

e. BANDWIDTH

The vast bandwidth provides easy transmission of data and as the visible light is license free it is free to use.

f. LOW COST

As this technology consists of very few components so it is cheaper comparatively.

g. FREQUENCY

Radio waves have lower frequency i.e. longer wavelengths due to which it consumes a lot more time when multiple devices get connected

5. LIMITATION OF Li-Fi

- The major drawback of this technology is that the waves can't pass through objects or penetrate through walls which results in data loss while transmission of data. If the receiver is unknowingly blocked anywhere then the signal gets immediately cut off.
- Another major cause regarding reliability and network in the path of transmission is the interruption of external sources that maybe sunlight, normal bulbs or any interruption. This interference causes disturbance in the communication as Li-Fi works on direct line of sight.
- Harald Hass says that the need of Wi-Fi is still required as light bulbs can't be available everywhere. We can't have a light bulb that renders high speed data to the moving object or provide data in any remote area that have many obstacles like walls, huts and so on.[3]

6. COMPARISION BETWEEN Li-Fi and Wi-Fi

As the problem defined above that due to the heavy traffic and connection of many devices only to a single router the speed of Wi-Fi degrades. To justify the need of Li-Fi, a comparison is below with some base points to distinguish: [12, 13]

a) Capacity

The data is transmitted through radio waves having a limited bandwidth which is also highly expensive. With this fast moving world the new technologies are developing like 3G,4G due to which we are running out of spectrum.

Whereas in comparison with radio waves bandwidth, the light waves bandwidth are 1000 times wider than it which provides a broader spectrum for data transmission.

b) Congestion

As more and more devices gets involve for data transmission through Wi-Fi, the complication increases and therefore it results in the degradation of speed which shows dull performance of this technology. But talking about Li-Fi , as there is direct line of sight the availability of LED light is everywhere so there is no problem in the number of devices getting connected.

c) Security

Wi-Fi uses radio waves which can pass through any object. These waves can be intercepted and can be used which is a security issue whereas Li-Fi doesn't offer this opportunity to the intruder as the light waves cannot penetrate through objects.

d) Availability

Due to the radio waves that are used in Wi-Fi, the cell phones are restricted in some areas like aircraft and petrol pumps whereas there are no such restrictions with the light waves. It is available in any area.

e) Speed

According to the standard 802.11a, Wi-Fi provides communication rate of 54mbps which can be extended with the available techniques up to 1gbps.

While Prof.Harald Hass has already exhibited 3gbps on a single colour. The speed can be extended upto 9gbps if there will be full colour i.e. RGB on a single LED. These LED lights offer a lot more potential for wireless connectivity.

7. AREAS WHERE Li-Fi IS A NEED

Li-Fi technology can be implemented in those areas where it is a need. The areas where the radio waves are restricted as it is harmful to humans and those areas where radio waves doesn't work because it creates hazardous impact, light waves can be a superman in those areas. Pointing out some areas where this technology is a necessity: [3]

1. Petrol pumps and petrochemical plants : Radio waves are strictly restricted
2. Education system: For high speed so that multiple devices can access it at a time.
3. Aircraft : Radio waves doesn't work here
4. Underwater applications for military operations: For secret operations to be carried out with no fear of getting snapped.
5. Street lamps : For free access
6. Hospitals: For medical purpose as radiation can be dangerous to the patients.
7. Traffic System: To control traffic.

8. FUTURE AND FURTHER ADVANCEMENT

Though the speed offered is 3gbps on a single LED with a single colour, but future enhancement can be made in the speed by using full colour i.e. the mixture of red, blue and green in a single LED or different LEDs with different colour which will provide a variance in light's frequency. Along with it, the mixture of RGB will give the opportunity to alter the frequency of light with each frequency encoding a different data channel. Moreover, an array of LED lights can be used for parallel transmission of information which will result in rapid transmission with less or no traffic because of its extreme high speed [9]. Adding up more yes we are ready for Li-Fi as all the constituents are available and all the procedure just it has to be put together and needs to be implemented. Li-Fi is the future technology for next generation.

Figure 7. Multicolour LED bulbs [16]

9. CONCLUSION

With the rapid increase in technologies and development of new devices that require connectivity is in the need of high speed with efficiency and effectiveness which Wi-Fi technology's sluggish performance is making it down. Li-Fi technology appears as an alternate solution to the problems faced with radio waves by providing higher frequency, high speed and proper transmission of data with security. It can't be said as a replacement of Wi-Fi technology as need of Wi-Fi is still there but the areas where Wi-Fi is not supported, this new technology will overcome those limitations and will act as a dissolvent to it.

10. REFERENCES

[1]http://deweydigest.com/dewey/wp-content/uploads/2013/11/LED-Bulb

[2]http://www.ermt.net/docs/papers/Volume_3/3_March 2014

[3]http://www.ijarcsse.com/docs/papers/Volume_3/11_November2013/V3I11-0434.pdf

[4]http://newtecharticles.com/site-content/uploads/2013/04/working-of-li-fi-technology.png

[5]http://www.academia.edu/8404489/Li-Fi_Light_Fidelity_The_future_technology_In_Wireless_communication

[6]http://www.ijarcsse.com/docs/papers/Volume_5/6_June2015/V5I6-0175.pdf

[7]http://sunpartnertechnologies.com/wp content/uploads/2015/12/img_LIFI_02.png

[8]http://www.ijcttjournal.org/Volume4/issue-4/IJCTT-V4I4P195.pdf

[9]http://www.ermt.net/docs/papers/Volume_3/3_March 2014/V3N3-181.pdf

[10]https://pulsesensor.files.wordpress.com/2011/07/photodiode-irled.jpg

[11]http://www.ijarcsse.com/docs/papers/Volume_5/6_June2015/V5I6-0175.pdf

[12]http://www.ijcta.com/documents/volumes/vol5issue1/ijcta2014050121.pdf

[13]http://purelifi.com/wp-content/uploads/2013/09/Shedding-Light-On-LiFi.pdf

[14]http://i724.photobucket.com/albums/ww245/MohsinShah11/em_spect.jpg

[15]http://www.intechopen.com/source/html/19222/media/image14.png

[16]http://s3.amazonaws.com/digitaltrends-uploads-prod/2013/10/led-rgb-lights.jpg

International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 88 - 91, 2016, ISSN:- 2319–8656

Question Level based Opinion Generation in Web based Interactive Systems

Rajimol R
Department of Computer Science and Engineering
Mangalam College of Engineering
Kottayam, India

Vinodh P Vijayan
Department of Computer Science and Engineering
Mangalam College of Engineering
Kottayam, India

Abstract: The greatest challenge in opinion generation system is to understand the meaning of the question asked and answering the question correctly based on the knowledge level of users. The solution for this is the question level based opinion generation system, which generate opinions by understanding the in-depth meaning of the question. Opinions are extracted from reviews on different websites. System uses lexical based algorithm for finding the semantic relation between words and thus obtains the keyword used to retrieve the opinions. System also uses transformation based classification to learn the system. Transformation classification is based on rules and is ends up when no more transformation is possible for the data.

Keywords: Opinion generation system, Lexical based algorithm, Machine learning, Transformation based classification, Keyword identification.

1. INTRODUCTION

Opinion generated for multiple levels of users are a challenging issue in an interactive opinion generation system. The question level based opinion generation system helps to generate opinions with intelligence. Some of the related works generates opinions directly from the set of reviews. But here the system generates opinions according to the level of question asked by the user. It uses lexical driven algorithm to find the semantic relation between the sentences and also make use of knowledge bases to make the system intelligent by learning.

The system collects reviews from different sites as similar to other opinion generation system but it also learns the system in order to answering different levels of questions. It also perform sentiment analysis and ranking to the collected reviews [1]. System performs parsing technique in order to identify the keyword that strongly represents the question. Then it performs depth value calculation. Based on the value, system retrieves related opinions and presents it to the user.

System first analyzes the question asked by the user. Then it performs appropriate keyword identification and depth calculation. This function can be accomplished by using the lexical driven algorithm. Lexical analysis classifies the sentence in to tokens or group of characters, which helps to identify the keywords easily from a sentence.

After question analysis and keyword identification, the system generates appropriate opinions for the user.

System uses transformation based classification for learning the system. Thus it helps to make the system intelligent. Usually the transformation classification starts with simple solutions and then its process like a cycle. It ends up when no more transformation is possible.

2. RELATED WORKS

Automatic expressive opinion generation system helps to generate expressive opinions rather than simple and common opinions [1]. This system generates expressive opinions for enjoyable conversation. It takes opinions from reviews and ranked them. It also performs sentiment analysis to classify the opinions into positive and negative opinions. Uniqueness of generated opinion is based on adjective frequency.

Topic relevance based opinion generation [2] focuses on the problem of searching opinions over general ideas. It applies ranking to documents which contains individual opinions. From the ranked documents opinions are retrieved.

Micro opinion generation is an unsupervised approach [3]. It deals with generating concise summary of opinions with maximum of 2 to 5 words. Since the opinions provided as summaries or small content, it helps the user to easily understand the generated opinions. Major idea is to use existing words in original text to compose meaningful summaries.

OPINE is a novel method for mining reviews in order to build a model with product features. It is an unsupervised information retrieval system and it uses relaxation labeling as the key concept [4]. It extracts the noun phrases from reviews and retains the nouns whose frequency greater than threshold value.

The talent to detect expected or valuable things is called serendipity. Wikipedia forms a graph like structure consists of article nodes and category nodes [5] called triplets. This uses supervised data for processing.

MONEA is efficient development platform architecture for multi functional robots. Firstly, it embodies the meta architecture for networked-robots. Secondly, it provides some development models. Finally, it doesn't require heavy weight middleware [6].

Speech based interactive information system is basically a question answering system. System works in two modes [7]. When user asks a question, it switches to retrieval/QA mode and generate answers. Otherwise it works in system recommendation mode.

Schema is a multi party interaction human robot [9]. System consists of multi user tracking and fusion module, multi party dialogue manager and virtual human and robot control module.

Opinion mining deals with extracting required data or information from large dataset, it can be a web. Sentiment analysis classifies the opinions into positive or negative polarities [8]

3. OPINION GENERATION SYSTEM

Automatic expressive opinion generation system is used to produce enjoyable conversations rather than simple or common sentences. In order to make deeper knowledge about the question, the system uses lexical driven algorithm and machine learning concepts in it. Opinion generation is a classic Natural Language Processing (NLP) problem. The opinion generation system also needs to be intelligent in order to realize the question and to generate suitable opinions for the question.

3.1 Classes, processing and context of questions

Question classes can be of different types such as factoid typed questions, non-factoid typed questions and some questions may need deeper knowledge in order to answer it. Question processing deals with the understanding of the questions or semantics of the questions given by the user. It also needs to understand the context of the questions before processing it.

3.2 Opinion extraction and formulation

Opinion extraction is basically depending on the level of question asked by the user. Simple extraction may be enough for certain questions. It may want the partial answers to be extracted from various sources and combine them. At the same time, also needs to make the results of the opinion generation system to be as natural as possible.

System uses a Transformation-Based Learning (TBL) algorithm to induce rules from the training data. TBL is a technique used to learn the system by providing set of rules that converts simple solutions into all possible set. Rules are expanding by using rule sets. The algorithm greedily selects the rule that reduces the error rate the most.

The system uses a parser, which shows the structure of the English Sentence and also performs deeper logical analysis. It can be used for relation extraction, where semantic relationships between the words are extracted. It is also responsible for keyword extraction, which extracts the important

4. ARCHITECTURE

The system receives user questions as input. Then it performs question and topic analysis in order to produce appropriate answer for the question.

System uses pattern based extraction of words for deeper knowledge about the question with the help of lexical driven algorithm and transformation based learning.

It also uses certain hypothesis to generate rules and to make the system more intelligent. Finally the system generates appropriate opinion for the question.

System can also perform additional opinion generation by using the reviews and the uniqueness of the opinions. These can be determined by the adjective frequency of the sentence as in automatic opinion generation systems [1].

Knowledge bases and data sources are used for generating opinions. Knowledge base is trained by using some set of rules, so that the system can give intelligent answers for different level of questions.

Architecture mainly consists of three modules: question analysis module, keyword identification and depth calculation module and opinion generation module.

4.1 Question analysis module

In the question analysis module, system analyzes the question asked by the user. Then it splits the question into different parts in order to identify the keywords in the questions.

Question analysis module also identifies whether the question belongs to factoid or non factoid type[1].

4.2 Keyword identification module

Keyword identification and depth calculation module first analyze the words in the question. Then by using lexical driven algorithm, it finds the semantic relation between the words and thereby identify the keywords from the question. Then it performs the depth calculation of the corresponding keyword and stored this value in the database for opinion retrieval.

4.3 Opinion generation module

Opinion generation module first search the opinions for corresponding questions in database and knowledge base based on the depth value of the keywords, So that the system can give opinions for different level of questions. It also uses transformation based classification for training or learning the system. Based on the depth of question, system generates appropriate opinions and output it to the user.

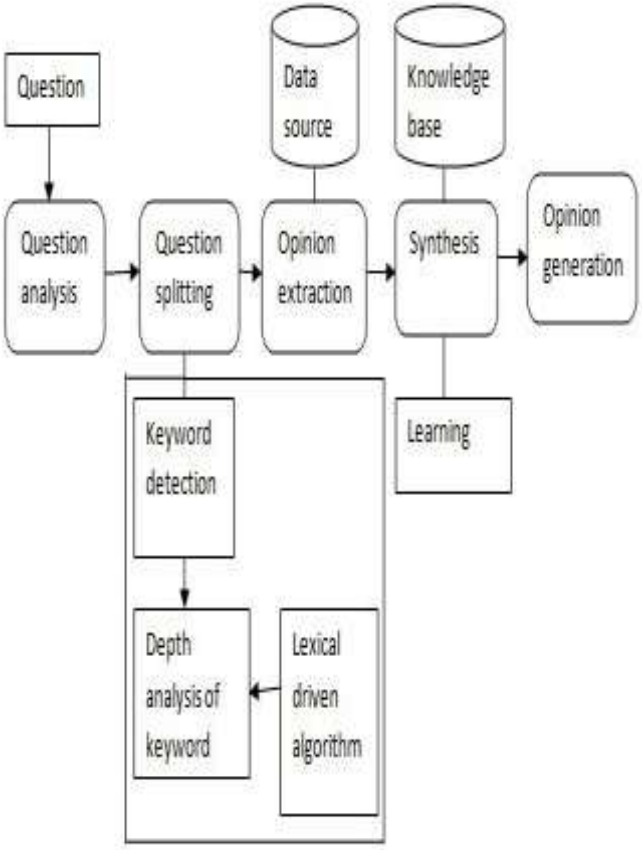


Figure 1. Architecture of question level based opinion generation system.

5. EXPERIMENTAL RESULTS

Experiments are conducted on Intel Core i3 processor with CPU of 2.40GHz. Data are extracted from different review sites. 14000 sentences are extracted from 20000 reviews. Comments to posts are not considered. To avoid mistakes done by sentence tokenizer, the length of the sentence must be < 200 characters.

Opinion generation systems are meant for generating opinions to users for their question. But most of the systems don't care about the satisfaction of the user. Satisfaction of the user can be

measured in many ways. Most preferable way is to measure the acceptance of opinion and response time of opinions.
The most important consideration for question level based opinion generation system is that whether they capable of providing opinions for complex or intelligent question also.
Question level based opinion generation system shows better performance in terms of acceptance of sentence and response time.

5.1 Acceptance of opinions

Most of the opinion generation system faces a problem for providing acceptable opinions that requires some knowledge. The problem is how to find the exact meaning of the question. This system solves this problem by identifying the keywords from the question. It also calculates the depth value for the keywords, so the system can easily understand the meaning and can provide more acceptable opinions.
Question level based opinion generation system produces more acceptable sentences. It uses learning techniques in order to retrieve and update the knowledge. So the system can produce more appropriate and context relevant sentence as output

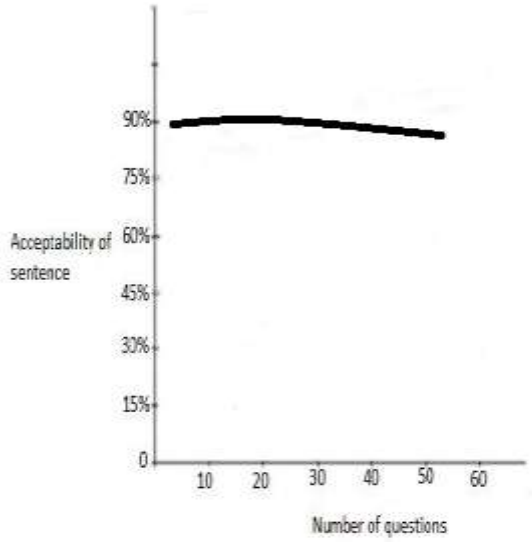


Figure 2: Acceptability of opinion

For measuring the performance of acceptability, plot a graph with number of questions in x-axis and acceptability of sentence in y-axis. Number of questions asked by the user can vary from 0 to any number. System shows better performance in any type of question, that is for factoid and non factoid typed questions.

System also checks whether the opinions given to the user is satisfied or not. It also provides better acceptable sentence even for intelligent questions.

Thus the system provides more acceptable opinions for any type of questions and any number of questions.

5.2 Response time

The response time of the system is directly related to the complexity of the sentence. As level of question becomes difficult, it may take more time to generate the opinions. Question level based opinion generation system shows better results even for complex questions.
In general opinion generation systems, it feels difficulty to provide opinions within specified time. So this may reduces the efficiency of the system and thus affect the users satisfaction.

These problems are happened due to the difficulty to understand the meaning of the question and also due to the unavailability of the data.
But in question level based opinion generation system, it uses data source as well as knowledge base for generating appropriate opinions. Knowledge base is periodically updated by using the transformation based classification. This classification or learning continues until no more selection is required.
Thus the opinions corresponding to any type of questions is available all the time. So the system gives response within few seconds even for the questions that needs intelligence.

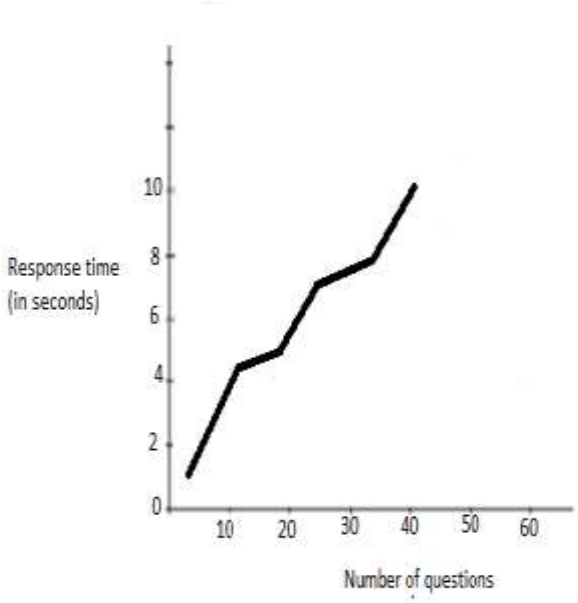


Figure 3: Response time of opinions

The response time of opinions can be measured by plotting number of questions on x-axis and response time in seconds on y-axis of a graph.

Here the number of questions can be varied from 0 to any number and to any level of question. The response time is plotted in seconds.

The obtained graph shows that the system gives answers for any number of questions within a short period of time (within seconds). System provides opinions within seconds even for different level of questions (for factoid, non factoid or complex questions).

Thus the question level based opinion generation system shows better performance in case of acceptability of opinions and response time of opinions. System also provides more accurate results as output.

6. CONCLUSION AND FUTURE SCOPE

Question level based opinion generation system is a type of web based opinion generation system which answers to a user query based on the level of question. The system shows an improved level of intelligence by understanding the depth of question and answering it correctly to the expected level of user. System is tested on a varying level of sample questions normally asked by the users. The generated opinions will be

helpful for the user to obtain suitable knowledge. It also performs depth value calculation of words in order to find the semantic relation between the words. This opinion generation system generates meaningful opinions by processing the meaning of the question and related data.
In future, it can generate audio based opinions as output.

7. REFERENCES

[1] Yoichi Matsuyama, Akihiro Saito, Shinya Fujie, and Tetsunori Kobayashi,"Automatic Expressive Opinion Sentence for Enjoyable Conversational Systems",in IEEE/ACM transactions on audio,speech and language processing, 2015 vol 23.

[2] Min Zhang , Xingyao Ye ," A Generation Model to Unify Topic Relevance and Lexicon-based Sentiment for Opinion Retrieval", in State key lab of Intelligent Tech.& Sys,2008.

[3] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A.Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, and J. Prager *et al.*,"Building Watson: An overview of the DeepQA project," *AI Mag.*, vol.31, no. 3, pp. 59–79, 2010.

[4] Ana-Maria Popescu and Oren Etzioni, "Extracting Product Features and Opinions from Reviews", Department of Computer Science and Engineering, University of Washington.

[5] Y. Noda, Y. Kiyota, and H. Nakagawa, "Discovering serendipitous information from Wikipedia by using its network structure," in *Proc.ICWSM*, 2010.

[6] T. Nakano, S. Fujie, and T. Kobayashi, "Monea: Message-oriented networked robot architecture," in Proc. IEEE Int. Conf. IEEE Robotics Autom. (ICRA), 2006, pp. 194–199.

[7] T. Misu and T. Kawahara, "Speech-based interactive information guidance system using question-answering technique," in *Proc. IEEE* Int.Conf. Acoust., Speech, Signal Process. (ICASSP '07), 2007, vol. 4, pp.IV–145.

[8] B. Pang and L. Lee, "Opinion mining and sentiment analysis,"*Foundat.Trends Inf. Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[9] Y. Matsuyama, K. Hosoya, H. Taniyama, H. Tsuboi, S. Fujie, and T.Kobayashi, "Schema: Multi-party interaction-oriented humanoid robot," in ACM SIGGRAPH ASIA Art Gallery Emerging Technol.:Adapt., 2009, pp. 82–82, ACM.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 95 - 98, 2016, ISSN:- 2319–8656

Co-Extracting Opinions from Online Reviews

Beema K S
Department of Computer Science and Engineering
Mangalam college of Engineering
Kottayam,India

Mitha Rachel Jose
Department of Computer Science and Engineering
Mangalam college of Engineering
Kottayam,India

Abstract: Exclusion of opinion targets and words from online reviews is an important and challenging task in opinion mining. The opinion mining is the use of natural language processing, text analysis and computational process to identify and recover the subjective information in source materials. This paper propose a Supervised word alignment model, which identifying the opinion relation. Rather than this paper focused on topical relation, in which to extract the relevant information or features only from a particular online reviews. It is based on feature extraction algorithm to identify the potential features. Finally the items are ranked based on the frequency of positive and negative reviews. Compared to previous methods, our model captures opinion relation and feature extraction more precisely. One of the most advantages that our model obtain better precision because of supervised alignment model. In addition, an opinion relation graph is used to refer the relationship between opinion targets and opinion words.

Keywords: ***Opinion mining, Sentiment Analysis, Topical Relation, Opinion Target, Opinion Words***

1. INTRODUCTION

Growth of web 2.0 huge number of user generated data is present on web as blogs, reviews, , comments etc. This data involve user's opinions beliefs, sentiment towards particular product, topic, event, news etc. An opinion mining refers to the use of natural language processing to extract the subjective information from source materials. Opinion mining includes opinion feature which is used to specify an attributes of an entity on which consumers state their views and opinions.

Others opinions can be crucial when it's time to make a judgment or choose among numerous opinions. Sentiment analysis is the computational study of people's emotions. Given a set of documents *D* that contain or sentiments about an object, opinion mining aims to extract attributes and means of the object that have been commented on in each document $d \in D$ and to evaluate the comments are positive, negative or neutral. This fascinating problem is increasingly important in business and society. It has loads of research challenges but promises approaching helpful to anyone interested in opinion analysis and social media analysis. Humans are objective creatures and opinions are significant Being able to interact with customers, has many advantages for information systems.

Textual information in the world can be broadly classified into two main categories, subjective and objectives. Facts are objective statements about entities and events in the web. It also makes it difficult for the producer of the product to keep track and to supervise customer opinions. For the manufacturer has an additional difficulties because lots of commercial sites may sell the similar product and the manufacturer normally produces many kinds of product. To extract the opinions from online reviews, it is unsatisfactory to obtain the overall sentiment about a particular product. That is an opinion has a positive and negative orientation. For Example:

"A stunning design and a big boost to Core i7 but poor battery life "

At this point an opinion about the laptop consisting positive opinion as "*stunning design and big boost to core i7"* and negative opinion *as "poor battery life"*

An opinion target is the object about which users express their opinion typically noun or noun phrase, in the above example *design.corei7,* and *battery* are the three opinion target. An opinion word is defined as the words that are used to express the users opinions. In the above example *stunning, big boost* and *poor* are the opinion words.

Rather than sentiment analysis and feature extraction proposed method mainly focused on Topical relation. That is extracting the current interest or relevance or pertaining or dealing with matters of current or local interest. This means that

2. RELATED WORKS

Lots of studies have paying attention on the task of opinion target and opinion word extraction[1], [2], [5] ,[6]. General textual inspection uses part of speech (POS) information (for example, nouns, adjectives, adverbs, and verbs) as a basic form of word-sense description. Some adjectives are good indicators of emotion and guide feature assortment to categorize the sentiment. Also, selected phrases elected by pre-specified POS patterns, usually including an adjective or adverb, help detect sentiments.

Pang and Lee [8] presented survey on sentiment analysis and opinion mining. So as toward survey they explained opinion oriented information right of entry, challenges, opinion categorization and summarization. Many researchers used machine learning methods for emotion examination [3] [4] [7] that involve guidance of classifier on datasets and use the skilled model for new document classification. Some authors optional another method such as dictionary of word lexicons [6].

Qiu et al. (2009,2011) proposed a *Double Propagation* technique [5] to explain a domain sentiment lexicon and an view target set iteratively. They exploited direct relations between words to extract opinion targets and belief words iteratively. The main limitation of Qiu's technique is that the patterns based on dependency parsing tree may introduce many noises for the great corpora (Zhang et al. 2010).

Oppressed syntax information [6] to extract opinion targets, and calculated some syntactic patterns to capture the opinion relations among words. The experimental consequences showed that their technique performed better than that of [5].

3. SYSTEM ARCHITECTURE

The opinion mining refers to the use of natural language processing, text analysis and computational linguistics to recognize and take out subjective information in basis resources. Opinion mining is generally useful to reviews and social media for a diversity of applications, ranging from marketing to customer service. As of the customer viewpoint, bearing in mind others opinions before purchasing a product is a common performance extended before the survival of Internet.

Rather than feature classification we focused on Topical Relations. In Topical Relation [1] extract the relevant features only from a particular product from online reviews. We first classify the sentences as opinions or facts and then we will examine only the subjective sentence thus improving performance. Also, we would add a smart crawler component so that all the relevant information from various web pages in a website is automatically crawled and extracted upon providing a URL and certain conditions. We determine the relationship between opinion targets and opinion words. We take all nouns are opinion targets and all adjectives are opinion words. An Opinion relation graph is used to refer the relation between opinion targets and opinion words. To model this process, construct a bipartite graph. A bipartite graph (or bigraph) [1] , whose vertices can be divided into two disjoint sets U and V (that is, U and V are each independent sets) such as opinion targets and opinion words that every edge connects a vertex in U to one in V.

The opinion mining tasks can be widely categorized based on the level at which it is done with the various levels being namely,

a. The document level

b. The sentence level

c. The feature level.

a. The document level

At the document level sentiment categorization of documents into positives and negatives. Which is done with the assumption made that each document focuses on a particular object and contains opinion from a single opinion holder.

b. The sentence level

At the sentence level, recognition of opinionated sentences amongst the reviews is done by classifying data into objective and subjective. Subsequently, sentiment classification of the sentences is done moving each sentence into positive, negative based on Naive bayes classifier.

c. The feature level.

At the feature level, diversity of tasks that are looked for identifying and extracting features from view. After that determining whether the opinions on the features are positive. At last grouping feature synonyms and producing a feature-based opinion summary of multiple reviews/text.

In Topical Relation extract the relevant features only from a particular product from online reviews. We first classify the sentences as subjective (opinions) or objective (facts) and then we will analyze only the subjective sentences thereby improving performance.. Also, we would add a smart crawler component so that all the relevant information from various web pages in a website is automatically crawled and extracted upon providing a URL and certain conditions.

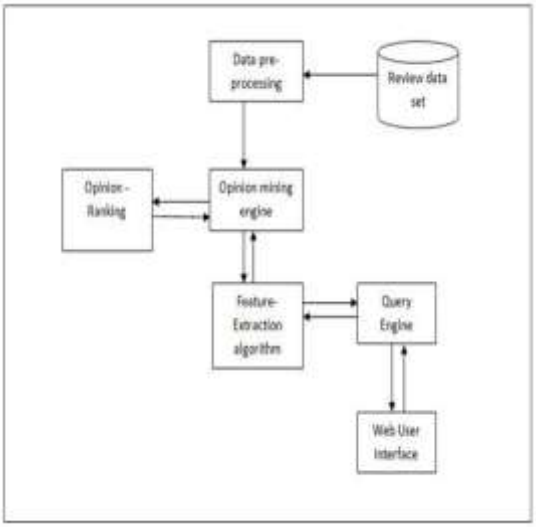


Figure.1 System Architecture

3.1 The Feature Extraction Algorithm

The algorithm wished to recognize potential features is called the Feature Extraction algorithm [9]. The idea behind the algorithm is that the nouns for which customers express many number of opinions are most likely to be the important and unique features than those for which users don't state such opinions. This algorithm takes an input is the list of adjectives which are used to express opinions. Pre-processing of words together with removal of stop words. Each and every sentences are parsed using Stanford Parser, then assign a Parts Of Speech (POS) tags to English words based on the context in which they appear.

.

4. EXPERIMENTAL EVALUATION

Consumers can place reviews on web communities, blogs, twitters, product's web site and these comments are called user generated contents. So huge number of data available freely in various websites. We experimented with different reviews on data set in order to measure several parameters of our system. More specifically, we performed three different sets of experiments. In the first line of experiments, we evaluated the performance of our opinion-based feature extraction algorithm, as compared to simple word count algorithms. Naive bayes classifier is used to correctly classifying reviews as positive or negative.

We conducted our experiments using the customer reviews of electronics products such as Iron Box ,Mobile Phone, Trimmers etc. The reviews are collected from the e-commerce site like Snapdeal.com and ebay.com. For each review, download the first reviews say 100 or 200 . Then Feature Extraction Algorithm is used to extract the product features and

also separate the opinion targets and the opinion words based on the word alignment model. By identifying the opinion targets we can detect whether the sentence is positive or negative. By using topical relation, we can extract the relevant information or features only from a particular product from reviews.

TABLE 1 Review Data Set

Data Items	Reviews
Iron Box	This is a very good and light weight press, very good and attractive look.
Mobile Phone	High performance, Poor Battery life ,High Resolution, etc.
Trimmers	Stunning Design, awesome product, It is worth to money and good performance,etc.

We select the review datasets for Iron Box, mobile phone, Trimmers. Reviews are first segmented into sentence, next sentence are tokenized using standard NLP tool. By analyzing the datasets we can separate the opinion targets. Measuring the performance of the product based on the positive and negative opinion targets. In Fig 2, shows the performance evaluation of Digital camera, Mobile phone and laptops on various years. In Fig .3 Topical Relation, we can detect the performance of a particular features from particular products.

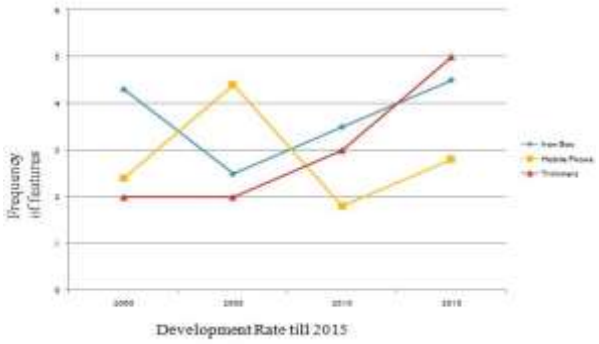


Figure .2 Performances of Products

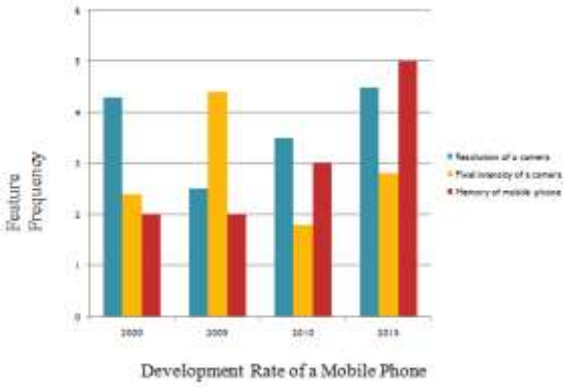


Figure.3 Product Improvement Based on Topical Relation

We can rank the products based on the customer reviews. By considering the Mobile phone, have positive and negative opinions. Fig:4 and Fig:5 showing the positive and negative ranking respectively.

There have lot of positive recommendations of the value of reviews for ecommerce, that the case doesn't really need to be made anymore, Fairly simply, user reviews increase conversions. They can eliminate any doubts possible customers may have about a product. In Fig: 4, shows that positive ranking of the mobile phone and Trimmer. Here 75% of reviewers say that the positive opinion about the memory of phone and 18% of pixel intensity and remaining reviews about the resolution of the phone. In case of Trimmer, 55% of reviewers say that positive opinion about the attraction of trimmer and 33.3% of weight and 22% is about the battery life of a trimmer.

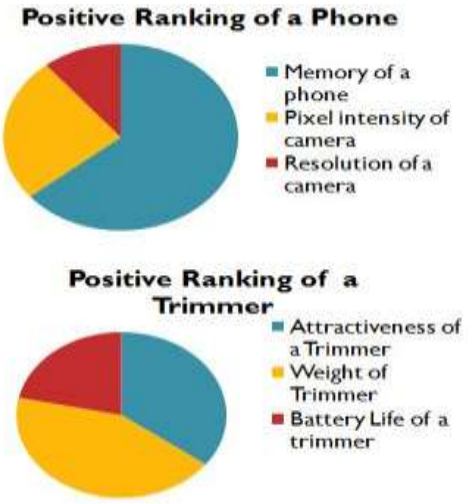


Figure 4. Positive Ranking

When purchasing items online, reading customer reviews is a suitable way to get a real account of other people's opinions of the product. Negative reviews that are set by a politeness-factor can actually help sell the item. In Fig:5, shows that negative ranking of the mobile phone. Here 25 of peoples say that the negative opinion about the memory and 76% of about pixel intensity and the remaining about the resolution of the phone. In case of Trimmer, 25% of reviewers say that negative opinion about the attraction of trimmer and 7% of weight and 68% is about the battery life of a trimmer.

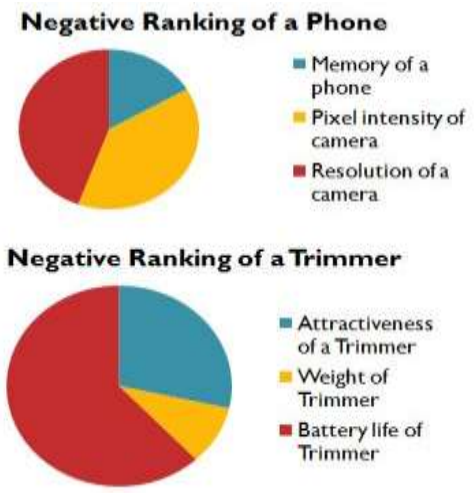


Figure.5 Negative Ranking of Mobile Phone

5. CONCLUSIONS

Opinions are the unique type of information which is different from facts. Joint information has spread all through the Web, particularly in areas connected to everyday life, like e commerce. Despite significant progress, however, opinion mining and sentiment analysis finding their own voice as new fields. This paper propose a Supervised word alignment model, which identifying the opinion relation. Rather than this paper focused on topical relation, in which to extract the relevant information or features only from a particular online reviews. Finally the items are ranked based on the frequency of positive and negative reviews.. We first classify the sentence as objective or subjective and then we analyze the adjectives or nouns thereby improving the performance. Compared to previous methods, our model captures opinion relation and feature extraction more precisely.

6. REFERENCES

[1] Kang Liu, Liheng Xu, and Jun Zhao, "Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model", *IEEE Trans . Knowledge and data Engineering*, Vol. 27, No. 3, March 2015

[2] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc. 10th ACM SIGKDD Int. Conf. Knowledge. Discovery Data Mining, Seattle, WA, USA, 2004

[3] A.Mukherjee and B. Liu, "Modeling review comments," *in Proc.50th Annual. Meeting Assoc. comput. linguistics*, Jeju, Korea, Jul.2012

[4] K. Liu, L. Xu, and J. Zhao, "Opinion target extraction using word based translation model," *In Proc. Joint* Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. *Learn.*, Jeju, Korea, Jul.2012

[5] Y. Wu, Q. Zhang, X. Huang, and L.Wu, "Phrase dependency parsing for Opinion mining," *in Proc. Conf* .Empirical methods Natural. Lang. *Process, S*ingapore, 2009

[6] M. Hu and B. Liu, "Mining opinion features in customer reviews," *in Proc.*19th Nat. Conf. Artif. Intell., San *Jose*, CA, USA, 2004

[7] B. Wang and H. Wang, "Bootstrapping both product features and opinion words from Chinese customer reviews with cross inducing," *in Proc.*3rd Int. Joint Conf. Natural Lang. *Process.*, Hyderabad, India, 2008

[8] Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis", *Foundations* and Trends in Information Retrieval Vol. 2, Nos. 1–2 (2008).

[9] L. Zhang, B. Liu, S. H. Lim, and E. O'Brien-Strain, "Extracting and ranking product features in opinion documents," in Proc. 23th Int.Conf. Comput. Linguistics, Beijing, China, 2010.

[10] J. Zhu, H. Wang, B. K. Tsou, and M. Zhu, "Multi-aspect opinion polling from textual reviews," in Proc. 18th ACM Conf. Inf. Knowl.Manage., Hong Kong, 2009, pp. 1799–1802.

[11] Z. Hai, K. Chang, J.-J. Kim, and C. C. Yang, "Identifying features in opinion mining via intrinsic and extrinsic domain relevance," IEEE Trans. Knowledge Data Eng., vol. 26, no. 3, p. 623–634, 2014.

[12] K. Liu, H. L. Xu, Y. Liu, and J. Zhao, "Opinion target extraction using partially-supervised word alignment model," in Proc. 23rd Int. Joint Conf. Artif. Intell., Beijing, China, 2013.

[13] A.M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in Proc. Conf. Human Lang. Technol. Empirical Methods Natural Lang. Process., Vancouver, BC, Canada, 2005.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 99 - 103, 2016, ISSN:- 2319–8656

Distributed Digital Artifacts on the Semantic Web

Susan P Kurian
Department of Computer Science and Engineering
Mangalam college of Engineering
Kottayam, India

Vishnu S Sekhar
Department of Computer Science and Engineering
Mangalam college of Engineering
Kottayam, India

Abstract: Distributed digital artifacts incorporate cryptographic hash values to URI called trusty URIs in a distributed environment building good in quality, verifiable and unchangeable web resources to prevent the rising man in the middle attack. The greatest challenge of a centralized system is that it gives users no possibility to check whether data have been modified and the communication is limited to a single server. As a solution for this, is the distributed digital artifact system, where resources are distributed among different domains to enable inter-domain communication. Due to the emerging developments in web, attacks have increased rapidly, among which man in the middle attack (MIMA) is a serious issue, where user security is at its threat. This work tries to prevent MIMA to an extent, by providing self reference and trusty URIs even when presented in a distributed environment. Any manipulation to the data is efficiently identified and any further access to that data is blocked by informing user that the uniform location has been changed. System uses self-reference to contain trusty URI for each resource, lineage algorithm for generating seed and SHA-512 hash generation algorithm to ensure security. It is implemented on the semantic web, which is an extension to the world wide web, using RDF (Resource Description Framework) to identify the resource. Hence the framework was developed to overcome existing challenges by making the digital artifacts on the semantic web distributed to enable communication between different domains across the network securely and thereby preventing MIMA.

Keywords: ***Digital artifacts, man in the middle attack(MIMA), semantic web, RDF, trusty URI***.

1. INTRODUCTION

With the ascend of credit cards, contactless payments & crypto currencies people have been predicting the end for physical money for nearly 60 years. Over the past decades, researchers have confirmed that there is only 9% of physical money with men and the rest is invested via internet, as technology has made work easier, which can be done from anywhere at any time. And here comes the relevance of this system to provide security for data in web, which is one among the greatest challenges currently raised. The solution for this is the distributed digital artifact system, which prevents the relevant man in the middle attack to an extent by ensuring verifiability and reliability.

The system consists of a coordinator process, to manage the domain which is assumed to be trusted. Seed generator is used to connect server in a domain which want to part of the semantic web publication, through which index of reference tree is built in multiple domain. Hash value will be calculated and Base 64 encoding is done. It is then published on the interface once the RDF encoding has been generated. On users request for service at the server, the server in turn connect to other server which has the required resource and the document is delivered to the client if the right access is satisfied followed by Base 64 decoding.

The rapid development of online payments, e-commerce sites, netbanking etc. made human work much easier, where they can do everything sitting at home on a button click. Due to these emerging developments website attacks have increased rapidly, where user security is at its threat. Among websites attacks man in the middle attack (MIMA) is a serious issue, where a malicious actor places himself into a conversation between 2 parties to access the information that they have send to each other. This work tries to prevent MIMA to an extent, by providing self reference and trusty URIs even when presented in a distributed environment. Any manipulation to the data is efficiently identified and any further access to that data is blocked by informing user that the uniform location has been changed.

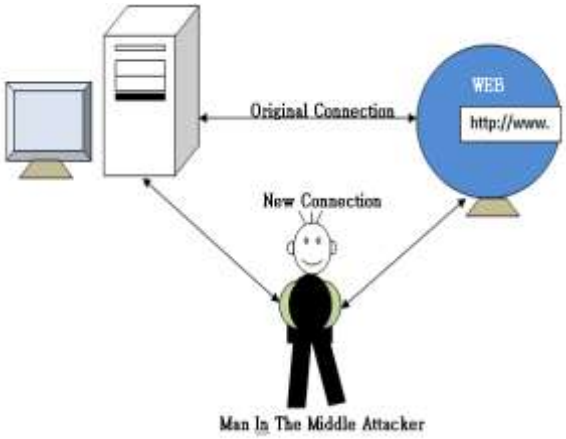


Fig.1 Man In the Middle Attack

2. RELATED WORKS

Lots of research are going on in the task of making digital artifacts on the web verifiable and reliable[1], [3], [5] ,[6]. In [1] authors suggest a module wise approach to make documents on the web correct, unmodified and always made available. The system makes use of trusty URIs[2] including hash values to identify modified input, which returns a totally changed value, even when slightly changed.

Nymble [2] is a system in which servers blacklist misbehaving users and blocks them. Websites use a seed for each nymble for blacklisting users, which in turn links future

nymbles from same user. It's a comprehensive credential system which maintains the privacy of blacklisted users.

Tobias Kuhn and Michel Dumontier in [3], a mechanism to incorporate cryptographic hash values in URIs was proposed . It was used to make the entire reference trees verifiable. The modular architecture used improves reliability and efficiency of tools.

Semantic web security and privacy system [4] deals with policy based security and privacy management on the Semantic Web. It supports protecting sensitive resources and information revelation. It describes policy, their interactions, specification, conflict detection and validation.

In [5] a decentralized approach to circulate, access and storing of data is considered. It propose a web based bottom-up process allowing researchers to publish, retrieve data in a reliable and trustworthy conduct.

Data lineage [7], [8] is used for checking data correctness. It describes data origin, how its extracted and its modification over time.

3. SYSTEM ARCHITECTURE

In centralized digital artifact system, when users request for service it will be fetched from the RDF stored in the central server and delivered. In semantic web which uses self-reference the verification occurs between a single central server and different URNs resulting in just a reference tree as output.

But in distributed digital artifact system, resources are distributed among different domains and each domain can communicate with each other. Here resources will not be stored in central server, rather will be distributed, and requests from users will be passed between different domains for processing. Here cross site verification is possible between different domains resulting in a complete forest as output. In distributed environment RDF is automatically generated which ensures efficiency of the system whereas in the other its externally generated which is a drawback consuming more time.

The system consists of different domains, which will be managed by coordinator process. Seed generator is used to generate a number for unique identification of multiple domains in a distributed environment. The resources will be distributed among multiple domains where they can communicate with each other. Hash value of a particular cited document will be calculated [8] and Base 64 encoding [1] is done. It can then be published on the interface once the RDF encoding has been generated. On users request for service at the server, the server in turn connect to other server which has the required resource and the document is delivered to the client if the right access is satisfied followed by Base 64 decoding.

Trusty URIs [1] will end with a hash value encoded in Bae64 notation, which can be a typical ASCII character (A-Z or a-z), any digit (0-9), a hyphen (-) or an underscore (_). All trusty URI will end with no less than 25 Base64 characters. The *artifact code*, whose first character represent type and second character version representing module identifiers which are followed by *data part*, which holds a hash part.

http://localhost:8080/r1.RA5AbXdpz5DcaYXCh9l3eI9ruBosiL5XDU3rxBbBaUO70

From the above example, localhost:8080/ represents self-references, resources that holds within, their own trusty URI. The whole thing that follows r1. is the artifact code. Its first character R recognize the module indicating its type and second character A specifies the version. The left behind 43 characters represent the real hash value.

The system consists of a server process which manage the various domains. The RDF generator process is responsible for generating metadata for the uploaded data in the session. Centralized and distributed storage of document/data is controlled by a storage process. Hashing make sure that data is integrated and encoding is employed for secure traversal of hash value. Client process verify the integrity of the document on the arrival at client side.

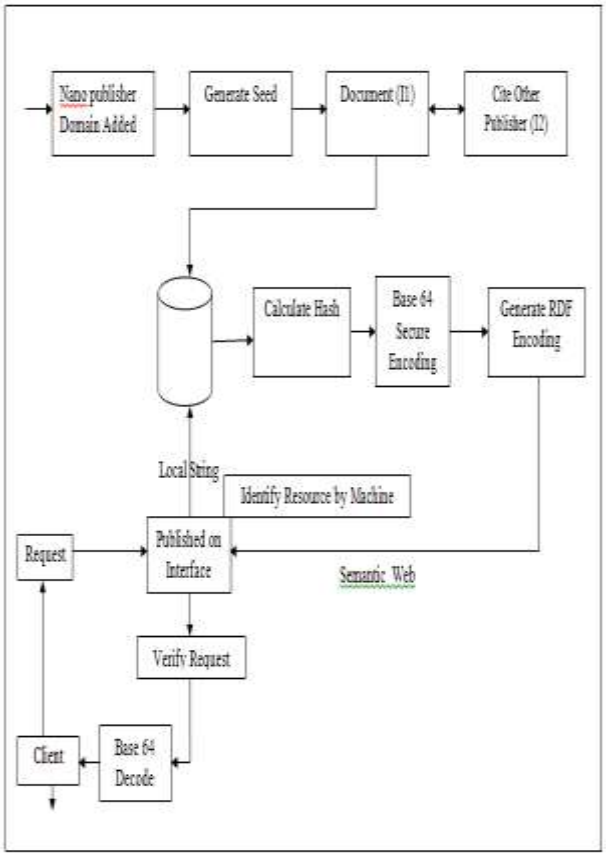


Fig.2 System Architecture

The modules of the proposed system can be broadly classified into the following namely,

a. Seed Generation
b. Distributed Communication
c. File Content Access
d. RDF Access
e. RDF Transferral

f. Client Request Processing

a. Seed Generation

Seed is a sequence of randomly generated number, providing unique id. A lineage algorithm [7] is employed. Whenever a domain is registered with the distributed system,its corresponding storage id is created which will be further used for its unique identification. Each domain will be linked to a seed, using which one domain will be connected to the other. Seed generator is used to connect server in a domain which want to part of the semantic web publication, through which index of reference tree is built in multiple domain.

On each user request, domain verifies seed to identify the site of the requested document to be delivered. Distributed network connects its different domains to each other where users can publish, retrieve and replicate documents distributed through the network.

b. Distributed Communication

Here each domain is free to communicate with each other, since the index of reference tree is built in multiple domain. A document can cite other publisher in a distributed environment. Each domain can post their publication to another domain or even to themselves. The domain to which posted can either approve or reject the document. But it requires no validation if posted to themselves. If approved its RDF [10] is automatically generated, and the document is published on the interface, accessible to all others across the network and if rejected its corresponding entry will be deleted. A domain himself acting as an attacker can sabotage the entrusted document given upon trust. But even presented in a distributed environment enabling inter-domain communication, the system ensures security to the document making digital artifacts on the web verified and trustworthy using *trusty URIs* and prevents *MIMA* attacks.

c. File Content Access

At FA, using SHA-512 hash generation algorithm [8] hash value is calculated, to which after appending two zero-bits are converted to Base64 notation generating trusty URN and complete trusty URI.

d. RDF Access

At RA, supports multiple graphs which works on RDF content. It allows self-references, resources that contain their own trusty URI. For Unicode characters a SHA-512 is generated in UTF-8 encoding, append two zero bits and is finally converted to Base64 notation.

e. RDF Transferral

At RB, trusty URI represents single RDF graph. Similar to RA, hash value is calculated for Unicode using SHA-512 in UTF-8 encoding and is transformed to Base64 notation.

f. Client Request Processing

The user request for service (finding, querying, filtering) at the server. The server in turn connect to other server which has the required resource. The connection requesting server has the hash index to verify that they are also in trusted domain. If the right access satisfied, Base64 decoding employed and the document is delivered to the client.. Any modification deny further access to that URL, returning an error message informing uniform location has been changed.

Integration or verification of trusty uri is made with the help of RDF meta data, which is machine understandable data. Whenever a domain uploaded the data, its corresponding hash value is included in the rdf tag with another metadata like seed, storage location etc. On browser's request for the document, the server respond with trusty uri which contain the hash value in the Base64 encoded form. When the document is loaded the client process calculate the hash value and perform matching function to do accept/reject.

3.1 The Seed Generation Algorithm

A lineage algorithm [7], [9] is used to generate a seed which is a randomly generated number for unique identification. Whenever a domain is registered with the distributed system, its corresponding storage id is created which will be further used for its unique identification. Each domain will be linked to a seed using which one domain will be connected to other. On each user request, domain verifies seed to identify the location of the requested content to be delivered. Every first post in each seed will be treated as *parent seed* which will be followed by *child seeds*. Each user request will processed from parent seed to childrens. The parent seed is searched using bubble sort, with a complexity of *O(n)* whereas childrens use quick sort with *O(nlogn)* complexity. The search is completed with an overall complexity of *O(nlogn)* which improves the performance.

4. EXPERIMENTAL EVALUATION

Experiments are conducted on Intel Core i3 processor with CPU of 2.40GHz. In order to measure several parameters of the system different data sets were experimented.

Distributed digital artifact system shows high performance than other systems in terms of MIMA detection rate and MIMA prevention rate.

A. MIMA Detection Rate

Man in the middle attack is a type of cyber attack where a malicious actor tries to get information that two parties send to each other. Since humans totally dependent on the internet, MIMA attacks have tremendously increased and preventing them is very essential.

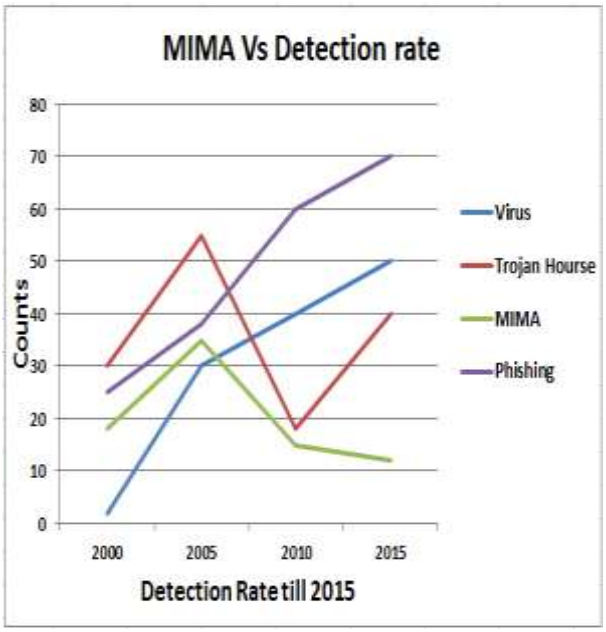


Fig.3 MIMA detection rate

Fig.3 shows that as years pass by the attacks rapidly increase. It illustrates a comparison between different attacks like virus, trojen hourse, phishing and man in the middle attacks and it shows that as years go man in the middle attack(MIMA) is on its hike and detecting MIMA is very difficult i.e, its detection rate has rapidly decreased which shows the relevance of this work.

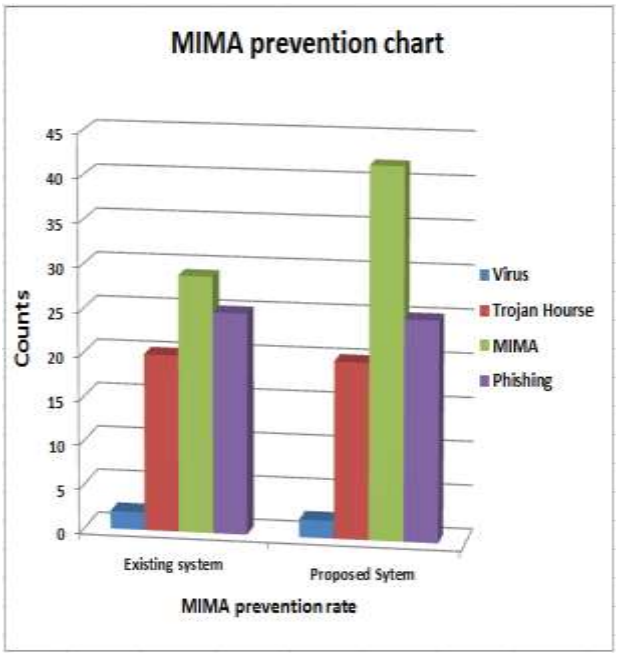


Fig.4 MIMA prevention rate

Fig.4 shows shows that MIMA is evidently prevented using this system compared to the existing. It offers security to the data in the semantic web using reference links. On the web, attacker constantly watches user practices and is vigilant of web applications. They always try to impose attacks on the network, by even slightly manipulating any content on the web. The user unknowing of the attack access the data which seems to be same as original and gets exposed to these attacks. Since fisching, online payments, e-commerce sites, netbanking etc. gained wide proliferation nowadays, these type of website attacks are very emerging and has become a serious issue.

5. CONCLUSION

The Distributed digital artifact system for MIMA is where resources are distributed among different domains and each domain can communicate with each other. Unlike centralized system, distributed system gives users possibility to check whether the data have been modified. The relevant man in the middle attack is prevented to an extent by ensuring verifiability and reliability. The system ensures that data published within the system interface cannot be accessed anywhere outside the system, with the use of reference trees providing security at an overall level. Any manipulation to the data is efficiently identified and any further access to that data is blocked by informing user that the uniform location has been changed. Here only man in the middle attack is considered. This can be extended to more attacks.

6. REFERENCES

[1] Tobias Kuhn and Michel Dumontier, “Making Digital Artifacts on the Web Verifiable and Reliable”, IEEE Transactions on Knowledge and Data Engineering, Vol NO 99 YEAR 2015

[2] Patrick P. Tsang, Apu Kapadia, Member, IEEE, Cory Cornelius, and Sean W. Smith, “Nymble: Blocking Misbehaving Users in Anonymizing Networks”, IEEE Transactions ON Dependable and Secure Computing

[3] Tobias Kuhn and Michel Dumontier, “Trusty URIs: Verifiable, Immutable, and Permanent Digital Artifacts for Linked Data”, in Proceedings of the 11th Extended Semantic Web Conference (ESWC 2014), ser.Lecture Notes in Computer Science. Springer, 2014

[4] N K Prasanna Anjaneyulu anna, Shaik Nazeer, “Semantic Web Security and Privacy”, Journal of Theoretical and Applied Information Technology

[5] Tobias Kuhn, Christine Chichester, Michael Krauthammer and Michel Dumontier, “Publishing without Publishers:a Decentralized Approach to Dissemination,Retrieval, and Archiving of Data”, arXiv preprint arXiv:1411.2749, 2014

[6] Momi Maity, Neha Verma, Rupali Wadikar, Sayali Shevkar, Prof. V.K. Bhusari, “Providing Security to Web Applications in Anonymizing Networks Using Nymble” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014 ISSN: 2277 128X

[7] Mingwu Zhang, Xiangyu Zhan, Sunil Prabhakar, “Cost Effective Forward Tracing Data Lineage”, Computer Science Technical Reports. Paper 1669, 2007

[8] S.FarrelL, C.Dannewitz, D.Kutscher, B.Ohlman, A.Keranen, P. Hallam-Baker, “Naming Things with

Hashes", Internet Engineering Task Force (IETF),April 2013

[9] Robert Ikeda and Jennifer Widom, "Data Lineage: A Survey", frmikeda,widomg@cs.stanford.edu

[10] C. Sayers and A. Karp, "Computing the digest of an RDF graph", Mobile and Media Systems Laboratory, HP Laboratories, Palo Alto, USA, Tech. Rep. HPL-2003-235(R.1), 2004.

[11] M. Bellare, O. Goldreich, and S. Goldwasser, "Incremental cryptography: The case of hashing and signing", in Advances in Cryptology — CRYPTO'94. Springer, 1994, pp. 216–233.

[12] R. D. Peng, "Reproducible research in computational science", vol. 334, no. 6060, p. 1226, 2011.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 104 - 109, 2016, ISSN:- 2319–8656

Educational Data Mining by Using Neural Network

Nitya Upadhyay
RITM
Lucknow, India

Abstract: At the present time, the amount of data in educational database is increasing day by day. These data enclose the concealed information that can lift the student's performance. Among all classification algorithms, decision tree is most algorithm. Decision tree provides the more correct and relevant results which can be beneficial in improvement of learning outcomes of a student. The ID3, C4.5 and CART decision tree algorithms are already implemented on the data of students to anticipate their accomplishment. All three classification algorithm have a limitation that they all are used only for small So, for large database we are using a new algorithm i.e. SPRINT which removes all the memory restriction and accuracy arrives in other algorithms. It is fast and scalable than others because it can be implemented in both serial and parallel fashion good data replacement and load balancing. In this paper, we are representing a new SPRINT decision tree algorithm which will used to solve the problems of classification in educational data system.

Key words: Educational Data mining, Classification, WEKA

1. INTRODUCTION:

Data mining is an emergent and rising area of research and development, both in academic as well as in business. It is also called knowledge discovery in database (KDD) and is an emerging methodology used in educational field to get the required data and to find the hidden relationships helpful in decision making. It is basically a process of analysing data from different perspectives and summarizing it into useful information (ramachandram, 2010). Now a day, large quantities of data is being accumulated. Data mining can be used in various applications like banking, telecommunication industry, DNA analysis, Retail industry etc.

Educational Data Mining: It is concerned with developing methods for exploring the unique types of data that come from educational database and by using data mining techniques; we can predict student's academic performance and their behaviour towards education (yadav, 2012). As we know, large amount of data is stored in educational database; data mining is the process of discovering interesting knowledge from these large amounts of data stored in database, data warehouse or other information repositories:

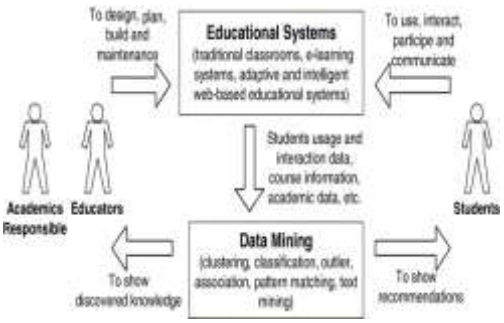


Figure 1.1- The cycle of applying data mining in educational system

Various algorithms and techniques are used for knowledge discovery from databases. These are as follows:-

- Classification
- Clustering
- Regression
- Artificial intelligence
- Neural networks
- Decision trees
- Genetic algorithm
- Association rules etc.

These techniques allow the users to analyse data from different dimensions, categorize it and summarized the relationship, identified during the mining process (yadav, 2012). **Classification** is one of the most useful data mining techniques used for performance improvement in education sector. It is based on predefined knowledge of the objects used in grouping similar data objects together (baradhwaj, 2011). Classification has been identified as an important problem in the emerging field of data mining. It maps data into predefined groups of classes (kumar, 2011). Classification is an important problem in data mining. It has been studied extensively by the machine learning community as a possible solution to the knowledge acquisition or knowledge extraction problem. The input to the classifier construction algorithm is a training set of records, each of which is tagged with a class label. A set of attribute values defined each record. Attributes with discrete domains are referred to as categorical, while those with ordered domains are referred to as numeric. The goal is to induce a model or description for each class in terms of the attribute. The model is then used by the classifier to classify future records whose classes are unknown.

2. LITERATURE SURVEY:

A number of data mining techniques have already been done on educational data mining to improve the performance of students like Regression, Genetic algorithm, Bays classification, k-means clustering, associate rules, prediction etc. Data mining techniques can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students.

Decision tree algorithm can be implemented in a serial or parallel fashion based on the volume of data, memory space

available on the computer resource and scalability of the algorithm. The C4.5, ID3, CART decision tree algorithms are already applied on the data of students to predict their performance. But these are useful for only that data set whose training data set is small. These algorithms are explained below:-

- **ID3**

Iterative Dichotomiser 3 is a decision tree algorithm introduced in 1986 by Quinlan Ross. It is based on Hunt's algorithm. ID3 uses information gain measure to choose the splitting attribute. It only accepts categorical attributes in building a tree model. It does not give accurate result when there is noise and it is serially implemented. Thus an intensive pre-processing of data is carried out before building a decision tree model with ID3 (verma, 2012). To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked.

- **C4.5**

It is an improvement of ID3 algorithm developed by Quilan Ross in 1993. It is based on Hunt's algorithm and also like ID3, it is serially implemented. Pruning takes place in C4.5 by replacing the internal node with a leaf node thereby reducing the error rate. It accepts both continuous and categorical attributes in building the decision tree. It has an enhanced method of tree pruning that reduces misclassification errors due to noise and too many details in the training data set.
Like ID3 the data is sorted at every node of the tree in order to determine the best splitting attribute. It uses gain ratio impurity method to evaluate the splitting attribute (baradhwaj, 2011).

- **CART**

It stands for **classification and regression trees** and was introduced by Breiman in 1984.It builds both classifications and regression trees. The classification tree construction by CART is based on binary splitting of the attributes. It is also based on Hunt's algorithm and can be implemented serially. It uses gini index splitting measure in selecting the splitting attribute. CART is unique from other Hunt's based algorithm as it is also use for regression analysis with the help of the regression trees (baradhwaj, 2011). The regression analysis feature is used in forecasting a dependent variable given a set of predictor variables over a given period of time. It uses many single-variable splitting criteria like gini index, sym gini etc and one multi-variable in determining the best split point and data is stored at every node to determine the best splitting point. The linear combination splitting criteria is used during regression analysis.

- **SLIQ**

It stands for **supervised learning in ques**. It was introduced by Mehta et al (1996). It is fast scalable decision tree algorithm that can be implemented in serial and parallel pattern. It is not based on HUNT'S Algorithm for decision tree classification. It partitions a training data set recursively using breadth-first greedy strategy that is integrated with pre-sorting technique during the tree building phase. The first technique used in SLIQ is to implement a scheme that eliminates the need to sort the data at each node of the decision tree. In building a decision tree model SLIQ handles both numeric and categorical attributes (Rissanem, 2010). Sorting of data is required to find the split for numeric attributes.

- **PUBLIC**

It stands for pruning and building integrated in classification. Public is a decision tree classifier that during the growing phase, first determines if a node will be pruned during the following pruning phase, and stops expanding such nodes. Hence, PUBLIC integrates the pruning phase into the building phase instead of performing them one after the other. Traditional decision tree classifiers such as ID3, C4.5 and CART generally construct a decision tree in two distinct phases. In the first building phase, a decision tree is first built by repeatedly scanning database, while in the second pruning phase, nodes in the built tree are pruned to improve accuracy and prevent over fitting (Rastogi, 2000).

- **Rainforest**

It provides a framework for fast decision tree constructions of large datasets. In this algorithm, we have a unifying framework for decision tree classifiers that separates the scalability aspects of algorithms for constructing a decision tree from the central features that determine the quality of the tree. This generic algorithm is easy to instantiate with specific algorithms from the literature (including C4.5, CART, CHAID, ID3 and extensions, SLIQ, Sprint and QUEST).
Rainforest is a general framework which is used to close the gap between the limitations to main memory datasets of algorithms in the machine learning and statistics literature and the scalability requirements of a data mining environment (Gehrke, 2010).

- **SPRINT algorithm**

It stands for **Scalable Parallelizable Induction of decision tree** algorithm. It was introduced by Shafer et al in 1996. It is fast, scalable decision tree classifier. It is not based on Hunt's algorithm in constructing the decision tree, rather it partitions the training data set recursively using breadth-first greedy technique until each partition belong to the same leaf node or class. It can be implemented in both serial and parallel pattern for good data placement and load balancing (baradhwaj, 2011).

Sprint algorithm is designed to be easily parallelized, allowing many processors to work together to build a single consistent model. This parallelization exhibits excellent scalability to the users.

It provides excellent speedup, size up and scale up properties. The combination of these properties or characteristics makes Sprint an ideal tool for data mining.

Algorithm:-

- Partition (data S)
- If (all points in S are of the same class) then
- Return;
- For each attribute A do evaluate splits on attribute A;
- Use best split found to partition S into S1 &S2;
- Partition (S1);
- Partition (S2);
- Initial call: partition (Training data)

There are 2 major issues that have critical performance implications in the tree-growth phase:

1. How to find split points that define node tests.
2. Having chosen a split point, how to partition the data.

It uses two data structure: attribute list and histogram which is not memory resident making sprint suitable for large data sets, thus it removes all the data memory restrictions on data.

It handles both continuous and categorical attributes. Data structures of SPRINT are explained below:-

Attribute list - SPRINT initially creates an attribute list for each attribute in the data. Entries in these lists, which we call attribute records, consist of an attribute value, a class label and the index of the record from which these values were obtained. Initial list for continuous attributes are sorted by attribute value once when first created.

- **Histograms** – Two histograms are associated with each decision-tree node that is under consideration for splitting. These histograms denoted as C_{below} which maintain data that has been processed and C_{above} which maintain data that hasn't been processed. Categorical attributes also have a histogram associated with a node.

However, only one histogram is needed and it contains the class distribution for each value of the given attribute. We call this histogram a count matrix. SPRINT has also been designed to be easily parallelized. Measurements of this parallel implementation on a shared-nothing IBM POWER parallel system SP2. SPRINT has excellent scale up, speedup and size up properties. The combination of these characteristics makes SPRINT an ideal tool for data mining (Shafer).

3. PRESENT WORK:

Decision tree classification algorithm can be implemented in a serial or parallel fashion based on the volume of data, memory space available on the computer resource and scalability of the algorithm. The main disadvantages of serial decision tree algorithm (ID3, C4.5 and CART) are low classification accuracy when the training data is large. This problem is solved by SPRINT decision tree algorithm. In serial implementation of SPRINT, the training data set is recursively partitioned using breadth-first technique.

In this research work, the dataset of 300 students have been taking from B.tech. (Mechanical Engineering) by considering the input parameters as: - name, reg. no., their open elective subject in 4th sem., midterm marks, end term marks, choice of Open elective subject, polling should be there? Yes or no, suggestion regarding polling: - if yes then why and if no then why? There are 9 OE subjects in B.tech. (ME) and because of limited sheets, most of the students do not get their own choice of subject. It could be effect on their performance in exam. So the output would come out to be how students are performing according to the choice of their preference.

Table 3.1: Example of attribute list of dataset

Marks	Grade	Rid
72	Good	0
83	Good	1
78	Good	2
91	Good	3
65	Average	4
52	Average	5
43	Average	6

Table 3.2: Dataset after applying pre-sorting

After Pre-sorting:

Objectives of Problem:

The objectives of the present investigation are framed so as to assist the low academic achievers in higher education and they are:-

- Identification of the choice of students in polling system which affects a student's Performance during academic career.
- Validation of the developed model for higher education students studying in various universities or institutions.
- Prediction of student's performance in their final exam.

In my proposed work, I am implementing SPRINT decision tree algorithm for improved classification accuracy and reduce misclassification errors and execution time. I have explained this algorithm and then apply serial implementation on it to find out the desired results. I am comparing it with other existing algorithms to find out which will be more efficient in terms of the accurately predicting the outcome of the student and time taken to derive the tree.

Data structures:

1. Attribute lists:

The initial list created from the testing set are associated with the root of the classification tree. As the tree is grown and nodes are split to create new children, the attribute lists belonging to each node are partitioned and associated with the children. The example of the attribute list is:

In sprint algorithm, Sorting of data is required to find the split for numeric attributes. It uses gini-splitting index for evaluate split. Sprint only sort data once at the beginning of the tree building phase by using different data structure. Each node has its own attribute list and to find the best split point for a node, we scan each of the node's attribute lists and evaluate splits based on that attribute.

Histogram: - Histograms are used to capture the class distribution of the attribute records at each node.

- **Performing the Split:**

  When the best split point has been found for a node, we execute the split by creating child nodes and dividing the attribute records between them. We can perform this by splitting the node's list into two as shown in figure 4. In our example, the attribute used in the winning split point is Marks. After this, we scan the list and apply the split test on it. Then we move the records to two new attribute list i.e. one for each new child. We have no test that we can apply to the attribute values for the remaining attribute lists of the node to decide how to divide the records. To solve this problem, we work with rids (Shafer).

As we partition the list of the splitting attribute i.e. marks, we insert rids of each record into a hash table to notify that the record was moved in which child. We can scan the list of the remaining attributes and probe the hash table after collected rids.
The output then tells us with which child to place the record. Splitting process is done in more than one step, if the hash table is large for memory.

- **Finding split points**:

During the process of making decision tree, the goal at each node is to determine the split point that best divides the dataset belonging to that node. The value of a split point depends upon how well it separates the classes. Many splitting have been proposed in the past to evaluate the goodness of the split. We need some function which can measure which questions provide the most balanced splitting. The information gain metric is such a function.

- **Measuring impurity**: - we have a data table that contains attributes and class of that attribute, we can measure homogeneity or heterogeneity of the table based on the classes. We can say that a table is pure or homogenous if it contains only a single class. If it contains several classes, then the table is impure or homogenous. There are so many indices to measure degree of impurity. Most common indices are entropy, gin index and classification error.

Entropy = $\sum_j -p_j \log p_j$

Entropy of a pure table is zero because the probability is 1 and log (1) = 0. Entropy reaches maximum value when all classes in the table have equal probability. For a data set S

Gini Index = $1 - \Sigma p_j^2$

In the above formula, Pj is the relative frequency of class j in S. If a split divides S into two subsets S1 and S2, the index of the divided data Gini split(S) is given by the following formula:
Gini split(S) = n1/n gini (S1) + n2/n gini (S2)

The advantage of this index is that its calculation requires only the distribution of the class values in each of the partitions. To find the best split point for a node, we scan each of the node's attribute lists and evaluate splits based on that attribute.
The attribute containing the split point with the lowest value for the Gini index is then used to split the node. Gini index of a pure table consist of single class is zero because the probability is 1 and $1-1^2=0$. Similar to entropy, gini index also reaches maximum, value when all classes in the table have equal probability.

Classification error = 1 – max {Pj}

Similar to entropy and Gini index, classification error index of a pure table is zero because the probability is 1 and 1-max (1) =0. The value of classification error index is always between 0 and 1. In fact the maximum Gini index for a given number of classes is always equal to the maximum of classification error index because for a number of classes n, we set probability is equal to p=1/N.

- **Splitting criteria**:

To determine the best attribute for a particular node in the tree we use the measure called information gain. The information gain, gain(S, A) of an attribute A, relative to a collection of examples S, is defined as

$$\text{Gain ratio} = \frac{\text{Gain(S, A)}}{\text{Split Information}}$$

The process of selecting a new attribute and partitioning the dataset is now repeated for each non terminal descendant node. Attributes that have been incorporated higher in the tree are excluded, so that any given attribute can appear at most once along any path.

4. RESULTS:

The proposed SPRINT decision tree algorithm is implemented in WEKA tool. It contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. In this, data can be imported in any format like CSV, Arff, binary etc. data can also read from URL or database using SQL. There are various models for classifiers like Naïve Bayes, Decision Trees etc. We have used classifiers for our experiment purpose. In this, the classify panel allows the user to apply classification SPRINT decision tree and other existing algorithms to the data set estimate the accuracy of the resulting model.

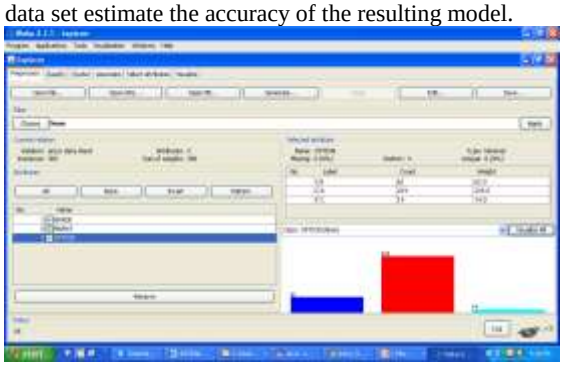


Figure 4.1: Preview after data set imported in Weka

In figure 4.1, Red colour implies that these attributes belong to option A, Blue colour implies that these attributes belong to option B and the green colour means that these attributes belong to option C.

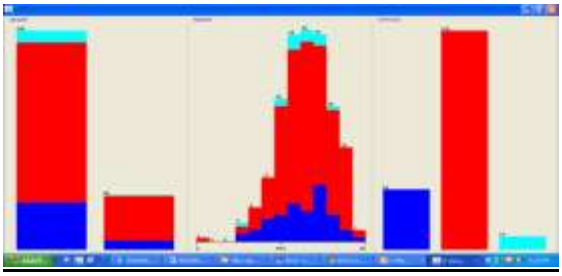


Figure 4.2: Visualizing all Attributes used in URL Classification

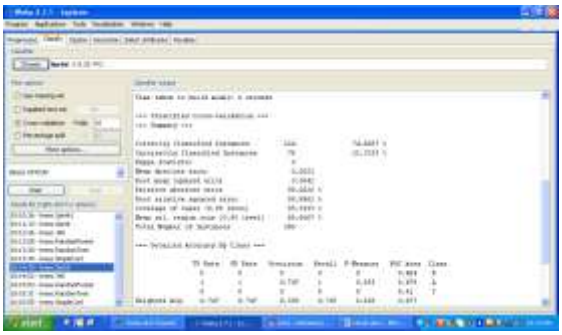


Figure 4.3: Classification by Sprint Decision tree

Figure 4.3 shows the comparison among all attributes on parameters like accuracy, true positive rate and false positive rate. The definitions of these terms are explained below:-

- **Accuracy:** The accuracy is the proportion of total number of predictions that were correct.
- **True Positive Rate:** The true positive rate (TP) is the proportion of examples which are classified as class x, among all examples which truly have class x, i.e. how much part of the class are captured. It is equivalent to recall.
- **False positive Rate:** The false positive rate (FN) is the proportion of examples which are classified as class X, but belong to a different class, among all examples which are not of class X.
- **Precision:** It is the proportion of examples which truly have class x among all those which are classified as class X.
- **F-Measure**: It is a combined measure for precision and recall defined by the following formula: -

  F-Measure = 2*Precision*Recall / (Precision + Recall)

4.1 COMPARISION:

The following table 1 shows the comparison between the working of different decision algorithms on the basis of different parameters.

Table 4.1:- Parameter Comparison of Decision tree algorithms

ALGORITMS	ID3 & C4.5	CART	SPRINT
Measure	Entropy info-gain	Gini diversity index	Gini Index
Procedure	Top-down decision tree construction	Constructs binary decision tree	Decision tree construction in a breadth first manner
Pruning	Pre-pruning using a single pass algorithm	Post pruning based on cost-complexity measure	Post-pruning based on MDL principle

4.2 OUTPUT

The three decision trees as examples of predictive models obtained from the data set of 300 students by three machine learning algorithms: C4.5 decision tree algorithm, random tree algorithm and the new SPRINT decision tree algorithm. Table 4.2 shows the simulation result of each algorithm. From this table, we can see that a Sprint algorithm has highest accuracy of 74.6667% compared to other algorithms. It also shows the time complexity in seconds of various classifiers to build the model for training data. By this experimental comparison, it is clear that Sprint is the best algorithm among four as it is more accurate and less time consuming.

Algorithm	Correctly classified instances	Incorrectly classified instances	Execution Time (sec)	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
SPRINT	74.6667%	25.3333%	0	0.2653%	0.3641%	99.2622%	99.9962%
J48	74.6667%	25.3333%	0.06	0.2653%	0.3641%	99.2622%	99.9962%
Random forest	67.666%	32.3333%	0.16	0.2714%	0.4123%	101.6039 %	113.2245%
Random tree	64.6667%	35.3333%	0.02	0.271%	0.4179%	101.4783 %	114.7643%

The result can vary according to the machine on which we are analysing our experiment. This is due to the specifications of the machine like processor, RAM, ROM and its operating system. However it will not affect the accuracy of the algorithm used.

5. CONCLUSION:

The efficiency of all the decision tree algorithms can be analysed based on their accuracy and time taken to derive the tree. The main disadvantages of serial decision tree algorithm (ID3, C4.5 and CART) are low classification accuracy when the training data is large. This problem is solved by SPRINT decision tree algorithm. SPRINT removes all the memory restriction and accuracy problem which comes in other existing algorithms. It is fast and scalable than others because it can be implemented in both serial and parallel fashion for good data placement and load balancing.

In this work, SPRINT decision tree algorithm has been applied on the dataset of 300 students for predicting their performance in exam on the basis of their choice in polling system. This result help us to find that the students who are opted their own choice of subject are giving better results than others.

6. REFERENCES:

[1] Brijsh Kumar bhardwaj and Saurabh Pal "Data mining: a prediction for performance improvement using classification", International journal of computer science an information security, vol. 9, no. 4, 2011.

[2] C.Romero and S.Ventra "Educational data mining: A survey from 1995 to 2005", 2006 Elsevier ltd. All rights reserved. www.elsevier.com/locate/eswa

[3] Dorina kabakchieva," Student performance prediction by using data mining classification algorithms", International journal of computer science and management research, Vol 1 issue 4 November 2012

[4] Devi Prasad bhukya and S. Ramachandram,"Decision tree induction- An Approach for data classification using AVL –Tree", International journal of computer and electrical engineering, Vol. 2, no. 4, August,2010.

[5] John shafer, Rakesh agrawal, Manish Mehta "SPRINT: A scalable parallel classifier for data mining" IBM Almaden Center, 650 Harry road, San Jose, CA 95120.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 110 - 114, 2016, ISSN:- 2319–8656

A Method for Sudanese Vehicle License Plates Detection and Extraction

Musab Bagabir,
College of Computer Science and Information Technology,
Sudan University for Science and Technology,
Khartoum, Sudan

Mohammed Elhafiz
College of Computer Science and Information Technology,
Sudan University for Science and Technology,
Khartoum, Sudan

Abstract: License Plate Detection and Extraction is an important phase of Vehicle License Plate Recognition systems, which has been an active research topic in the computer vision domain in order to identify vehicles by their license plates without direct human intervention. This paper presents a simple, fast and automatic License Plate Detection method for the current shape of Sudanese license plate. The proposed method involves several steps: green channel extraction, edge detection, regions of interest selection, dilation operation with especial structural element and connected component analysis. In order to analyze the performance and efficiency of the proposed method a data set for Sudanese vehicles has been created. Using this new data set, number of experiments has been carried out. Comparing with other countries license plate detection the achieved results is satisfactory.

Keywords: Green Channel Extraction; Edge Detection; Morphological Operation; Connected Component Analysis

1. INTRODUCTION

Vehicle License Plate Recognition (VLPR) systems is an important component for automating many control and surveillance systems, such as road traffic monitoring, parking lots, access control, highway electronic toll collection, red light violation enforcement, finding stolen cars and gathering traffic flow statistics [1]. Due to the differences of license plates in formats, styles, colors and size from one country to another, the field of VLPR and its applications has attracted many researchers in many countries to search and develop systems that solve these different problems. Therefore, so far, many methods have been proposed for VLPR depending on the country's license plate characteristics.

License Plate Detection (LPD) has been considered as the most important and essential phase of VLPR systems, which is directly influences the success and the accuracy of VLPR systems [3, 4]. For that, LPD requires more attention; moreover detecting a license plate on a complex background is a difficult task. Thus there are many factors should be considered in order to successfully detect and extract the license plate, for example: image quality, different plate sizes and designs, plate location and Background details and complexity [3].

According to Sarfraz et al the license plate extracted from the gray-scaled image by detecting vertical edges using Sobel edge detector, which uses a 3x3 mask, then filtering out unwanted regions by applying seed-filling algorithm [5]. The license plate region extracted by comparing the size ratio of the rectangular area between two vertical regions with the actual standard size ratio of the license plate. Alginahi uses the method reported by Sarfraz et al without filtering step in order to locate different License Plate types in shapes and size [6]. But Basalamah works depends on finding the black cross that centers Saudi Arabian plate, so an edge detector is applied to find the horizontal and vertical maps, then before median filter performed, the binary image is obtained by using the average value of pixels in each map as a threshold [7].

Abulgasem et al proposed Radial Basis Function Neural Network (RBF NN) to detect the Libyan license plate. First sobel edge detector applied, and then some morphological methods is used to thicken edges and remove unwanted edges. The remaining regions are detected and categorized into "plate" and "not plate" manually to train the RBF NN, which afterwards is used to detect the license plate automatically in other images during the testing phase [8].

Mousa presented an algorithm for Palestine LPD based on canny edge detector. This edge detection method is used to find image's edges based on local maxima of the gradient, which calculated by the derivative of a Gaussian filter [9]. Rasheed et al use canny edge detection operator and Hough lines to detect and extract the license plate [10].

The method proposed by Mohammad et al for LPD based on identifying the location of screws that hold the plates in place using pattern matching, plate aspect ratio (width to height ratio), and intensity levels. Then by applying coordinate system, the plate area is masked with respect to each screw position [11].

Shidore et al proposed LPD technique for Indian vehicles by using Sobel filter, morphological operations and connected component analysis [12]. Indian LPD mechanism developed by Davis et al, in which the gray-scaled image is converted to binary image by using adaptive thresholding. Then applied unwanted lines elimination algorithm based on 3x3 mask, which is moved throughout the image to identify the central pixel and testing the remaining 8 neighbor pixels. After that the vertical edges is extracted fallowed by highlighting the required regions technique. Once again unwanted lines elimination algorithm is used. As last step, the image is scanned for continuous black pixel in order to obtain the two diagonal corners of the license plate [13].

Deb et al proposed Sliding Concentric Window (SCW) based system to detect Korean license plate. After applying SCW on vehicle image, HSI color model is used for candidate region

color verification witch based on hue and intensity in HSI color model [14].

The remainder of this paper is structured as follows: Section 2, introduces the Sudanese vehicle license plate. Section 3, describes the proposed method. The experimental results are provided along with discussion in section 4. Section 5 concludes the paper.

2. SUDANESE CAR LICENSE PLATE

The Sudanese vehicles license plates are categorized in a number of types [2], that categorization was based on the differences of plates background color and characters color, Table 1. gives some information about these types.

Table 1. Sudanese Vehicles Plates Types

Type	Background Color	Characters Color
Private vehicles	White	Black
Commercials (Passenger)	White	Green
Commercials (Goods)	Black	White
Police	Blue	White
Government	Yellow	Black

The size of all plate types is 32 ×16 centimeters (see Figure 1). The plate has been divided into three regions; one region at the top part of the plate, which contains the name of the country "SUDAN" written in English and Arabic. The other two regions at the bottom part of the plate. They were separated by a silver metallic bar (Old Plates) Figure 1(a), or a vertical text "جمهورية السودان" (New Plates) Figure 2(b).The right bottom part consists of numerals (1 to 5 numbers) written in English and Arabic. Where the left bottom part consists of characters or a character and number written in English and Arabic, the characters are an abbreviation of Sudan states names, and the number to keep the sequence of the numbering. This study will focuses on the first type: private vehicles, as shown in Figure 1 (a) and (b).

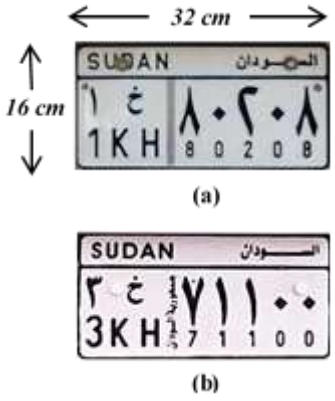


Figure 1. Sudanese License Plate (a) Plate with Silver Metallic Bar (b) Plate with Vertical Text

3. THE PROPOSED METHOD

The proposed method designed for Sudanese vehicle LPD. It composed of four of stages, including green channel and edge detecting, region of interest filtration, dilation and candidate regions detection and accurate plate detection/extraction, as shown in Figure 2.

The input of the method is the original image of the vehicle in RGB scale of size 2048×1536 pixels taken from real scene. The details of other stages are presented in the following subsections.

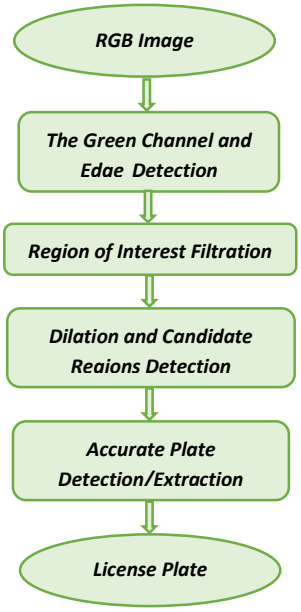


Figure 2. The Proposed Method

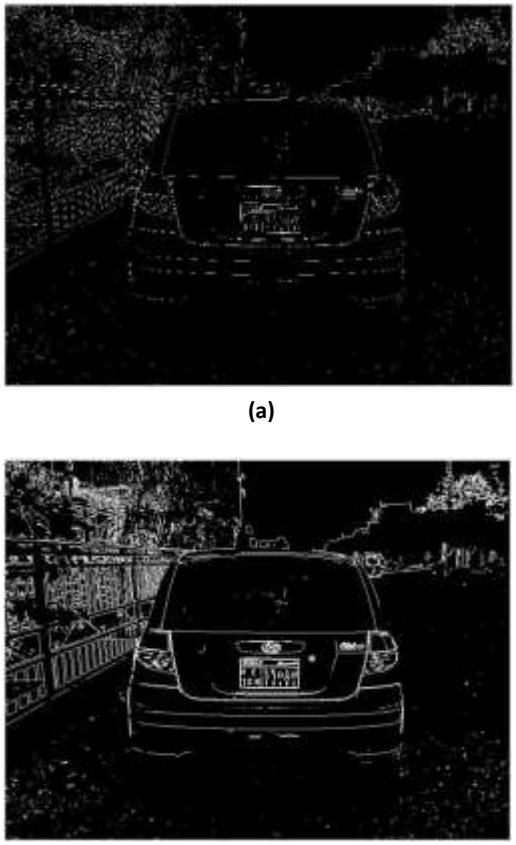
3.1 The Green Channel and Edge Detection

The RGB image consists of three channels red, green and blue, the value of each channel in the range 0- 256, whereas the gray scale image contains only one channel [15]. Thus, extraction of one channel will decrease the computational time, as well as the storage space. Experiments showed that the green channel provides sufficient contrast for the image, which in turn directly increases the efficiency of the proposed method than other channels or gray conversion. Thus, green channel extracted as in Figure 3. Afterwards, median filter is applied to remove noises like random occurrences of black and white pixels.

Figure 3. Green Channel Extraction

The next step is edge detection. Edges in images are areas with strong intensity contrasts, which represent a boundary between different regions. Detecting the edges of an image significantly reduces the amount of data and it helps in filtering out the useless information.

Primary investigation of this research shows that sobel edge detection has better results on Sudanese vehicles data set. The Sobel edge detector applies a 3x3 mask on the input image and gives the resultant binary image (see Figure 4(a)). Then, dilation operation with disk structural element is performed to thicken the edges, that is due to; edge's lines do not completely cover the region of interest. Figure 4(b) shows the resultant image.



(a)

(b)

Figure 4. (a) Edge Detection (b) Thicken Edge

3.2 Region of Interest Filtration

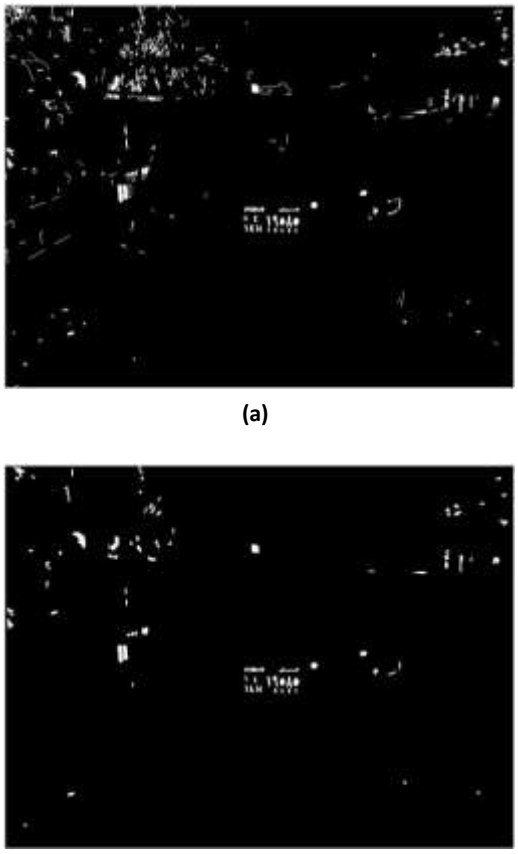
Filtration process is performed either to select regions that satisfy some particular features or eliminates unwanted regions on the image [16]. Filtration get the main aim of this step done, which is to obtain a filtered image has as possible all license plate contents except its boundaries. The step begins with removing the very small regions based on the number of white pixels of each one (Region Area). Afterwards, all regions in the resultant image were detected; their widths and heights used as features to select regions have a specific width and height. Although license plate contents are successfully well segmented, but there are some regions belong to the background are also selected, as shown in Figure 5 (a).

Thus, those unwanted regions should be removed or reduced. This is achieved by filling each region in the image and calculating their areas. Then, the region is selected if its area is greater or equal to A_{min} and less or equal to A_{max}, otherwise the region is removed, where:

A_{min}: Minimum Region Area

A_{max}: Maximum Region Area

A_{min} and A_{max} have been set during the experiments. Figure 5(b) shows the resultant image.



(a)

(b)

Figure 5. (a) License Plate Contents Selection (b) Remove Unwanted Regions

3.3 Dilation and Candidate Regions Detection

The aim of this stage is to obtain the candidate regions that might be the license plate region. It is based on the idea of merging the closed regions as well as removing unwanted regions. Therefore, the morphological operation dilation has been used with specific structuring element (SE) to expand the regions. When regions expand, the gaps reduced.

According to Sudanese car license plate layout as mentioned in section 2, the dilation operation is used three times with different SE values to merge each group of the license plate components separately. For instance, suppose that the license plate is divided into two parts vertically from the metallic bar, and then each part (Left/Right) contains three rows of characters as illustrated in Figure 6.

Figure 6. Visualization of How the Plate Vertically Divided

First performing dilation with SE value of size 30×15 pixel to merge the three rows of characters of each left and right part separately. Then removing any object contiguous to the border of the resultant image and others objects those their area less than the expected area of merged each two parts as in Figure 7(a). The purpose of second and third dilation is to join the two parts horizontally. This is achieved by using a special single row-SE to restrict the expansion of objects/regions along the horizontal direction only. After the second dilation, the bigger objects than a specific area value were selected, and then third dilation performed as shown in Figure 7(b).

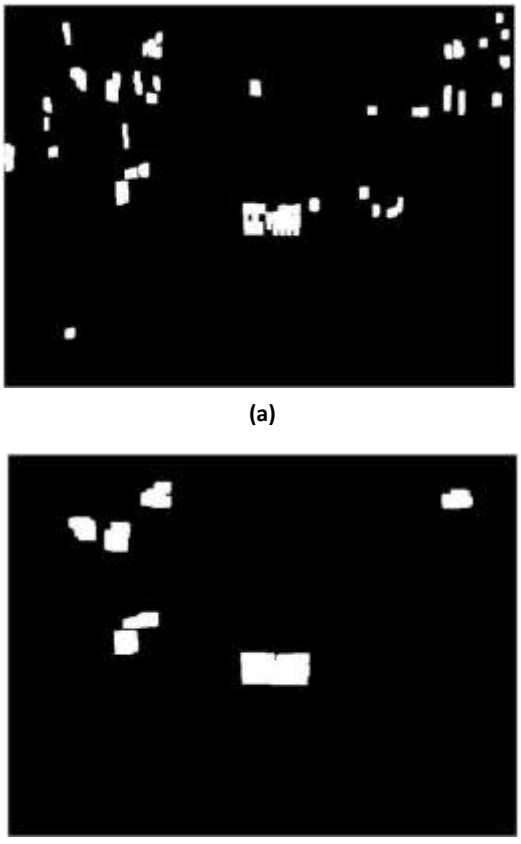


Figure 7. (a) First Dilation. (b) Second and Third Dilation

3.4 Accurate Plate Detection/Extraction

The license plate detection and extraction stage identifies the accurate region of the license plate. The connected components (objects/regions) analysis is performed to identify each object/region in the resultant binary image from previous stage. For each connected component in the image, some features are calculated in order to identify the license plate region. While each connected component surrounded by smallest rectangle as illustrated in Figure 8, those features are explained as follows:

a. Height to width ratio (aspect ratio).

b. The rectangle area (height × width).

c. The possible number of white pixels in the rectangle.

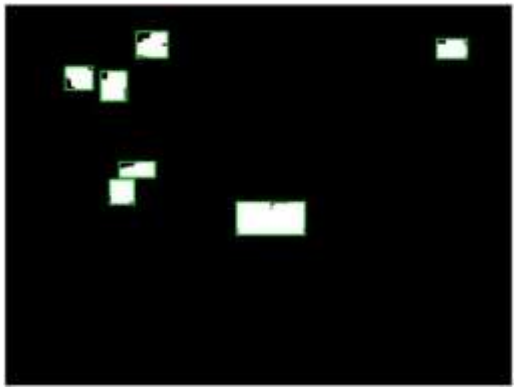


Figure 8. Connected Component are Surrounded by Rectangle

Afterwards, the accurate plate region is detected if the region position in the lower two thirds of the image. Figure 9 shows the result of this stage.

Figure 9. Extracted License Plate

4. EXPERIMENTS AND DISCUSSIONS

Experiments have been carried out under MATLAB R2013a (Version 8.1.0.604) environment to test the proposed method and to measure its accuracy. 200 color images with size of 2048 × 1536 pixels used for testing the method. Tested images have been captured in real scene, and the distance between the camera and the vehicle varied from 2 up to 3 meter.

The results of the Experiments are presented in Table 2. It shows the detection and extraction of the plate region accuracy is 98.5%.

Table 2. Experiments Results

Total Number Of Images	Successful Extraction	Unsuccessful Extraction
200	197	3
100%	98.5%	1.5%

On the other hand, the failure in detection and extraction can be due to damaged plates, the characters color changed to white, decorative items covered the plate and the reflection of the sun on the plate region as illustrated in Table 3.

Table 3. Failure in Detection and Extraction

Failed Detection	Comments
	License Plate characters color changed to white
	License Plate covered by decorative items
	Reflection of the sun on the plate region

5. CONCLUSION

The purpose of this paper is to presents a method for an automatic vehicle license plate detection and extraction. The proposed method is mainly designed for Sudanese license plate, according to the literature it is considered as first of its kind for Sudanese vehicle license plates.

According to the state of the art of vehicle license plate extraction, the proposed method is implemented through four stages: green channel and edge detecting, region of interest filtration, dilation and candidate regions detection and accurate plate detection/extraction.

The proposed method succeeds in detecting and extracting the plates efficiently and accurately with high rate percentage (98.5% for the given data set).

6. REFERENCES

[1] Hu, H., Zhang, Z., & Bai, Y. (2012). Car License Plate Location Based on Mathematical Morphology. In Recent Advances in Computer Science and Information Engineering. Springer Berlin Heidelberg: 415-420.

[2] Bagabir, M., Mariyam, S., Elhafiz, M. and Ahmed, A. (2015). Multi Objective Segmentation for Vehicle License Plate Detection with Immune-based Classifier: A General Framework. International Journal of Computer Applications Technology and Research 04 (04), 322 - 326.

[3] Du, S., Ibrahim, M., Shehata, M., & Badawy, W. (2013). Automatic license plate recognition (ALPR): A state-of-the-art review. Circuits and Systems for Video Technology, IEEE Transactions on, 23(2), 311-325.

[4] Patel, C., Shah, D., & Patel, A. (2013). Automatic Number Plate Recognition system (ANPR): A survey. International Journal of Computer Applications, 69(9), 21-33.

[5] Sarfraz, M., Ahmed, M. J., & Ghazi, S. (2003, July). Saudi Arabian License Plate Recognition system. In Geometric Modeling and Graphics, 2003. Proceedings. 2003 International Conference on (pp. 36-41). IEEE.

[6] Alginahi, Y. M. (2011). Automatic Arabic License Plate Recognition. International Journal of Computer and Electrical Engineering, 3(3), 454-460.

[7] Basalamah, S. (2013). Saudi License Plate Recognition. International Journal of Computer and Electrical Engineering, 5(1), 1.

[8] Abulgasem, N. A., Mohamad, D., & Mohamad Hashim, S. Z. (2011). Automatic License Plate Detection and Recognition Using Radial Basis Function Neural Network. International Journal of Computer Vision and Applications (IJCVA), 1(1).

[9] Mousa, A. (2012). Canny Edge-Detection Based Vehicle Plate Recognition. International Journal of Signal Processing, Image Processing & Pattern Recognition, 5(3).

[10] Rasheed, S., Naeem, A., & Ishaq, O. (2012). Automated Number Plate Recognition using hough lines and template matching. In Proceedings of the World Congress on Engineering and Computer Science (Vol. 1, pp. 24-26).

[11] Mohammad, K., Agaian, S., & Saleh, H. (2011, February). Practical Automatic Arabic License Plate Recognition System. In IS&T/SPIE Electronic Imaging (pp. 78810V-78810V). International Society for Optics and Photonics.

[12] Shidore, M. M., & Narote, S. P. (2011). Number Plate Recognition for Indian Vehicles. IJCSNS, 11(2), 143.

[13] Davis, A. M., Arunvinodh, C., & Arathy Menon, N. P. (2015, March). Automatic License Plate Detection Using Vertical Edge Detection Method. In Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on (pp. 1-6). IEEE.

[14] Deb, K., Khan, I., Saha, A., & Jo, K. H. (2012). An Efficient Method of Vehicle License Plate Recognition Based on Sliding Concentric Windows and Artificial Neural Network. Procedia Technology, 4, 812-819.

[15] Gonzalez, R. C., Woods, R. E., & Eddins, S. L. (2009). Digital Image processing using MATLAB®. United States: Gatesmark Publishing.

[16] Ibrahim, N. K., Kasmuri, E., Jalil, N. A., Norasikin, M. A., Salam, S., & Nawawi, M. R. M. (2013). License Plate Recognition (LPR): A Review with Experiments for Malaysia Case Study. The International Journal of Soft Computing and Software Engineering. (JSCSE), 3(3).