IJEST

ISSN: 0973-6255
Vol. 12 No.1
January - June 2018

Indian Journal of Engineering, Science, and Technology

A Refereed Research Journal

Published by

BANNARI AMMAN INSTITUTE OF TECHNOLOGY

(Autonomous Institution Affiliated to Anna University of Technology, Coimbatore - Approved by AICTE - Accredited by NBA and NAAC with "A" Grade)
Sathyamangalam - 638 401 Erode District Tamil Nadu India
Ph: 04295-226340 - 44 Fax: 04295-226666
www.bitsathy.ac.in E-mail: ijest@bitsathy.ac.in

BANNARI AMMAN INSTITUTE OF TECHNOLOGY
Stay Ahead

Indian Journal of Engineering, Science, and Technology

IJEST is a refereed research journal published half-yearly by Bannari Amman Institute of Technology. Responsibility for the contents rests upon the authors and not upon the IJEST. For copying or reprint permission, write to Copyright Department, IJEST, Bannari Amman Institute of Technology, Sathyamanagalam, Erode District - 638 401, Tamil Nadu, India.

Indian Journal of Engineering, Science, and Technology

ISSN: 0973-6255

Volume 12 Number 1 **January - June 2018**

CONTENTS

***Excerpt from the Proceeding of National Conferences**

***National Conference on Innovations in Information Technology (NCIIT-18) held at BIT 16-17 March 2018**
***National Conference on Challenges and Opportunities in Sensing Materials and Automation (COSMA'18) held at BIT 15-16 March 2018**

Optimization Technique for Effective Document Clustering

S.Thanmughi, A.Ramya Devi, N.PriyaDharshini and Mr.K.Thirukumar
Department of Computer Science and Engineering,
Dr. Mahalingam College of Engineering and Technology, Pollachi - 642 003, Tamil Nadu
E-mail: sudarthanmughi@gmail.com,thirukumar@drmcet.ac.in

Abstract

Document clustering is a process of automatic organization of documents into clusters so that documents within a cluster have high similarity in comparison to documents in other clusters. It has been studied intensively because of its wide applicability in various areas such as web mining, search engines, and information retrieval. It measures similarity between documents and groups similar documents together. It provides efficient representation and visualization of the documents, thus helps in easy navigation. A number of optimization algorithms are introduced to obtain the global optimal solutions. The optimization algorithms play an important role to obtain the best accurate results in the retrieved document. The information is growing normously and it is difficult and tedious task to retrieve the necessary information from that pool. The main area for retrieving relevant answers is called intelligent information retrieval. To achieve this, question and answering system is used.

This question and answering plays a major role in user query processing, information retrieval and extracting related information from the information pool.K-means Clustering is a partition clustering method in which each cluster is associated with a centroid (centre point). Clusters are formed based on centroid distance. A number of optimization algorithms is introduced to obtain the accurate and better results. Genetic Algorithm and Cuckoo Search are nature inspired met heuristic optimization algorithms. In this project, combination of Genetic Algorithm with Cuckoo Search is applied to the question and answering system. The proposed system uses PSO algorithm. The experimental results using PSO will be better in improving the accuracy in clustering the documents and optimizing the results.

1. INTRODUCTION

Question answering (QA) is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that automatically answer questions posed by humans in a natural language. A QA implementation, usually a computer program, may construct its answers by querying a structured database of knowledge or information, usually a knowledge base. More commonly, QA systems can pull answers from an unstructured collection of natural language documents. The first step in retrieving answers from the natural language documents using document clustering technique. Document clustering involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document clustering is generally considered to be a centralized process. Examples of document clustering include web document clustering for search users. The question processing has to analysis the query given by the user and retrieves the relevant documents from the web source repositories. The information extraction starts within the documents and proceeds with the extraction of accurate answers from the paragraph within the documents.

The extracted documents has various answers within itself. The more accurate document containing the answer is optimized using a number of algorithms. Recently, the number of optimization algorithms are introduced for obtain the global optimum solutions. Some of the optimization algorithms are Genetic Algorithm (GA), Ant Colony Optimization (ACO), Differential Evolution (DE) and Particle Swarm Optimization (PSO), Cuckoo Search (CS), Artificial Bee Colony (ABC), etc. The optimization algorithms are plays an important role in obtain the best accurate results in the retrieved answers.

Particle swarm optimization (PSO) is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. It solves a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical.

Formulae over the particle's position and velocity. Each particle's movement is influenced by its local best known position, but is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles. This is expected to move the swarm toward the best solutions.

In this paper, the PSO algorithm is applied for optimization of the clustered documents which contains the more accurate answers, from the available documents.

2. RELATED WORKS

This section[2] describes a strongly knowledge-based approach to question answering. As described in the following subsections, this approach relies on several types of knowledge. Among them, answer typing ("Qtargets"), semantic relationship matching, paraphrasing, and several additional heuristics all heavily rely on parsing, of both the question and all answer sentence candidates.

Nature inspired computing[4] is a technique that is inspired by nature. Nature has constantly served an inspiration for numerous scientific and technological developments. These nature inspired techniques are used to build numerous algorithms to solve optimization problems and find global optimal solution.

Consensus clustering [5] is finding a clustering solution in a set of clustering that is in the agreement with them. Consensus clustering, in literature, is also known as clustering aggregation, clustering ensembles, and clustering combinations. It can be considered as a meta-clustering technique where multiple clustering whether same or different, are combined to give an optimal clustering solution. This topic has attracted many studies in variety of areas. Given a set of m clustering Z1, Z2,..,Zm, consensus clustering tries to find a clustering Z that is in the least disagreement with the m clustering. Job scheduling problem [6,7] has a combinatorial optimization problem. Job scheduling is used in compound equipment manufacturing system for authenticating the performance of heuristic algorithms. Hence for scheduling problems it is too difficult to define a common frame work. For solving combinatorial optimization problem, an efficient algorithm is necessary. In this section, the proposal of hybrid algorithm for job scheduling which will combine the advantages of ACO and Cuckoo search.

Job scheduling [5,6] is a combinatorial problem. To solve a combinatorial problem, it is necessary to design an efficient algorithm. In this project, the proposal of an efficient hybrid algorithm that combines the advantage of both genetic and cuckoo search algorithm. This designed algorithm solve problem of job scheduling very effectively. There are N number of jobs and M number of machines. Each machine has its own order of execution. The main objective of proposed hybrid algorithm is to minimize the make span time. Make span time is the total time taken by

- Job should be a finite set.
- Machines should be a finite set.
- Every job must contain a series of operation that should be performed by machine.
- All jobs should be able to handle only one.

Particle Swarm Optimization [13] (PSO) is a self-adaptive global search based optimization technique introduced by Kennedy and Eberhart. The algorithm issimilar to other population-based algorithms like Genetic algorithms but, there is no direct re-combination of individuals of the population. Instead, it relies on the social behavior of the particles. In every generation, each particle adjusts its trajectory based on its best position (local best) and the position of the best particle (global best) of the entire population. This concept increases the stochastic nature of the particle and converge quickly to a global minima with a reasonable good solution.

Question Answering systems [8] are advanced search engines that can provide the least brief and the most complete answer to users instead of making them read a set of documents. QA systems are essential tools for dealing with the fast-growing global information. However, upgrading a search engine to a QA system is a complex and open-ended problem. Machine based human-like answering has been a dream that Artificial Intelligence (AI) scientists have been trying to achieve. Based on Russell and Norvig, AI field has four definition groups and one of them is based on the Turing test, which is about the ability of machines to communicate or answer like a human. Moreover, Arthur Samuel in his talk titled "AI: where it has been and where it is going" stated the main goal of AI and machine learning as: "to get machines to exhibit behavior, which if done by humans, would be assumed to involve the use of intelligence."

As the outcome of the survey and related works done by the various researchers, the advantages and

disadvantages of algorithms have been taken for consideration, the proposal of particle swarm optimization technique for the question answering system to make it as an intelligent interactive system.

3. EXISTING SYSTEM

The problem of mapping the user query to find the answer among the long list of documents, the search space is more complex and increases in response time. The system gets the query from the user in natural language through a interface, the question is analyzed using a POS-tagger for the question type such as evaluative question, hypothetical question, confirmative rhetorical question, non-factoid question from the learning model.

The learning model is trained with the set of question pairs of grammar arrangement of questions in a phrase tree format. The question is preprocessed for removing the stop word and stemming the words to extract the keywords from the question.

Keyword Extraction: KWN = U^{N}i=1Ext(Pos_noun(di))

The datasets are processed by grouping the documents with respect to the related domain context on the keyword basis. The semantic meaning of the keyword is tested using Wordnet Dictionary, an online dictionary with thesaurus.After the processing of the dataset documents with the finding of the word sense and word order, the knowledge base is built.

As a subsequent activity the related document to be ranked for the extraction, there is a chance of more than one document contain the keyword in group and then the normalized rank of document calculated using the formula:

Document rank = Rank of the document / No. of retrieved documents

After ranking the document, learning model maintains a list of documents with rank based on the query keyword along with the occurrence of the keyword in the list of the documents. The retrieved documents is taken for the extraction of related paragraph and sentence, the answer relevance score is calculated by using the minimum number of words between a keywords and candidate answer and string distance metrics. After obtaining the list of n sentences are reflecting as the population and apply to HGACSO algorithm for the answers.

4. PROPOSED SYSTEM

A basic variant of the PSO algorithm works by having a population (called a swarm) of candidatesolutions (called particles). These particles are moved around in the search-space according to a few simple formulae. The movements of the particles are guided by their own best known position in the search-space as well as the entire swarm's best known position. When improved positions are being discovered these will then come to guide the movements of the swarm. The process is repeated and by doing so it is hoped, but not guaranteed, that a satisfactory solution will eventually be discovered.

The procedure of the particle swarm optimization technique is given below:

. . . . . *i*= 1, ...,S do Initialize the particle's position with a uniformly distributed random vector:xi ~ . . **blo,bup**)

Initialize the particle's best known position to its initial position: pi ← xi

if*f*(**p**i) <*f*(**g**) **then**
update the swarm's best known position: **g** ← **p**i
Initialize the particle's velocity: **v**i ~ *U*(-|**bup**-**blo**|, |**bup**-**blo**|)
whilea termination criterion is not met **do**:
foreach particle*i*= 1, ...,*S* **do**

foreach dimension*d*= 1, ...,*n* **do** Pick random numbers: *r*p, *r*g ~ *U*(0,1)

Update the particle's velocity: **v**i,d ←
ω **v**i,d+ φp*r*p(**p**i,d-**x**i,d) + φg*r*g(**g**d-**x**i,d)

Update the particle's position: **x**i ← **x**i + **v**i

if*f*(**x**i) <*f*(**p**i) **then**

Update the particle's best known position: **p**i ← **x**i
if*f*(**p**i) <*f*(**g**) **then**

Update the swarm's best known position: **g** ← **p**i

The values **blo** and **bup** are respectively the lower and upper boundaries of the search-space. The termination criterion can be the number of iterations performed, or a solution where the adequate objective function value is found. The parameters щ, цр, and цg are selected by the practitioner and control the behavior and efficacy of the PSO method.

The following figure 1 depicts the architecture flow diagram of the proposed system.

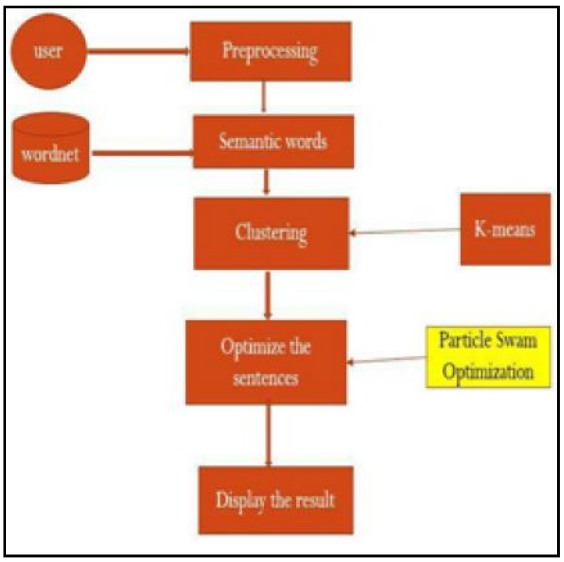


Fig.1 System Architecture of proposed system

The datasets are processed by grouping the documents with respect to the related domain context on the keyword basis. The semantic meaning of the keyword is tested using wordnet 3.0, an online dictionary with thesaurus. After the processing of the dataset documents with the finding of the word sense and word order, the knowledge base is built. The query keyword is matched with the documents in the knowledgebase and retrieves the most relevant document from the knowledgebase with the semantic similarity between them.

4.1 Preprocessing

The data sets are pre-processed and the user query is proceeded to be extract the root words.

- User enter query into the interface of the system.
- Process the query by removing the stop words and stemming of the words.
- Each word is tagged as noun, verb or adjective respectively using the tagger files.
- The semantic meaning of each word is given for the clustering purpose.

4.2 Clustering of Documents using K-means Algorithm

Document clustering refers to unsupervised classification. Classification of documents into groups(clusters) in such a way that the documents in a cluster are similar. Here, we cluster the documents using K-means algorithm. The K-means clustering algorithm is known to be efficient in clustering large data sets. K- user defined or pre-defined constant.

4.3 Optimizing using Particle Swarm Optimization

It optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. It solves a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulaeover the particle's position and velocity.

The dataset used in this experiment is the 20Newsgroup dataset. The dataset considered for training and testing the hybrid algorithm for ranking the data from the cluster. This is a very standard and popular dataset used for evaluation of many text applications, data mining methods, machine learning methods, etc.
Its details are as follows:
Number of unique documents = 18,828
Number of categories = 20
Number of unique words after removing the stopwords = 71,830

As a sample, the dataset taken contains 5 domain .Each domain consists of 40 documents with a total of 200 documents. Some of the attributes in the dataset are Business, Sports, Entertainment, Politics, Technology. The Datasets are processed by grouping the documents with respect to the related domain context on the keyword basics. The type of the Document is Text Document.

Table 1 20Newsgroup Dataset

Dataset	Domain	No. of Documents Used
20 Newsgroup	Business, Sports, Entertainments, Politics Technology	200

After the successful execution of the proposed system, we test our system with the test dataset to analysis with the help of 100 questions. The non-relevant document with different questions and non-relevant sentences are eliminated to enhance the speed up of the result to increase response time by the proposed PSO algorithm.

5. Evaluation Metric

The metrics used for the text mining techniques to verify the correctness of the results achieved based on the standard metrics like Precision, Recall, F-Score, Fallout and miss rates. The impact of features on the answer extractions with the considerations of the parameters analyzed are true positive, true negative, false positive, false negative, true and false positive rates.

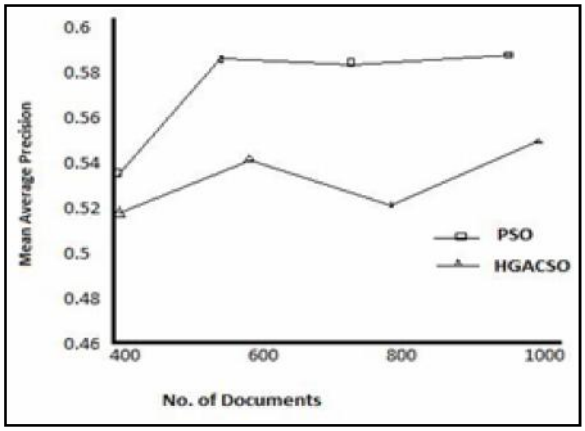
The precision is also called as positive predictive rate is the used to find the relevant sentences from the retrieved one by relevant sentences intersect with the retrieved sentences to the retrieved sentences.

$$\text{Precision} = \frac{\text{Relevant Sentences} \cap \text{Retrieved Sentences}}{\text{Retrieved Sentences}}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

wheretp is true positive and fp is false positive.

Experimental results show that by using information from the external corpora, PSO models produce imperative improvements on question pattern, document clustering based on domain context classification tasks and fallout & miss out rate are decreased, especially on datasets with few or short documents.



6. CONCLUSION AND FUTURE WORK

In this paper, particle swarm optimization is applied to the question and answering system. The proposed algorithm PSO is tested with the 20newsgroup datasetsThe results are compared with hybrid Genetic Algorithm and Cuckoo Search algorithms. The PSO algorithm outperforms compared to Genetic Algorithm and Cuckoo Search optimization algorithms. By using the PSO algorithm, the efficiency of the answer retrieval system is improved. The system provide the exact solution in less time, the future work is to analysis the user behavior and create a FAQ knowledge base for frequently and unanswered question.

REFERENCES

[1] Xin-She Yang and Suash Deb, CuckooSearch via Levy Flights", Proceedings of World Congress on Nature and Biologically Inspired Computing, 2009, pp. 210-214.

[2] J. Jeon, W. Croft and J. Lee, "Finding Semantically Similar Questions based on Their Answers", Proceedings of 28th AnnualInternational ACM SIGIR Conference on Research and Development in Information Retrieval, 2005, pp. 617-618, .

[3] IztokFister Jr., Xin-She Yang, IztokFister, Janez Brest and DusanFister, "A Brief Review of Nature-Inspired Algorithms for Optimization", Elektrotehni Ski Vestnik, Vol. 80, No. 3, 2013, pp.1-7.

[4] Nitisha Gupta and Sharad Sharma, "Nature-Inspired Techniques for Optimization: A Brief Review", International Journal of Advance Research in Science and Engineering, Vol.5, No.5, 2016, pp.36-44.

[5] MansafAlam and KishwarSadaf, "Web Search Result Clustering based on Cuckoo Search and Consensus Clustering", Indian Journal of Science and Technology, Vol.9, No.15, 2016, pp.1-18.

[6] R.G. BabuKartik and P. Dhavachelvan," Hybrid Algorithm by the advantage of ACO and Cuckoo Search for Job Scheduling", International Journal of Information Technology Convergence and Services, Vol.2, No.4, 2012, pp. 25-34.

[7] Satyendra Singh, JitendraKurmi and SudanshuPrakashTiwari, "A Hybrid Genetic and Cuckoo Search Algorithm forJob Scheduling", International Journal of Scientific and Research Publications, Vol. 5, No. 6, 2015, pp.1-4.

[9] OleksandrKolomiyets and Marie-Francine Moens, "A Survey on Question Answering Technology from Information RetrievalPerspective", Information Sciences, Vol.181, No.24, 2011, pp.5412-5434.

[11] Gunnar Schroder, Maik Thiele and Wolfgang Lehne, "Setting Goals and Choosing Metrics for Recommender System Evaluations", Proceedings of 5th ACM Conference on Dresden University of Technology Recommender Systems, 2011, pp.78-85.

[12] ImanKhodadi and Mohammad SanieeAbadeh, "Genetic Programming based feature Learning for Question Answering", Information Processing and Management, Vol.52, No.2, 2016, pp.340-357.

[13] M. Bhuvaneswari, S. Hariraman, B. Anantharaj and N. Balaji, "Nature Inspired algorithms: AReview ",International Journal of Emerging Technology in Computer Science and Electronics, Vol.12, No.1, 2014, pp. 21-28.

[14] IztokFister Jr., Xin-She Yang IztokFister,Janez Brest and DusanFister, "A Brief Review of Nature-Inspired Algorithms for Optimization", Elektrotehni SkiVestnik, Vol.80, No.3, 2013, pp.1-7.

[15] Pinar Civicioglu and ErkanBesdok, "A Conceptual Comparison of the Cuckoo Search, Particle Swarm Optimization, Differential Evolution and Artificial Bee Colony Algorithms", Artificial Intelligent Reviews, Vol.39, No.4, 2011, pp. 315-346.

[17] JohnH.Holland,"Adaptationin Natural and Artificial Systems: An Introductory Analysis with Applications to Biology,Control,andArtificial Intelligence", MIT Press, 1975.

A Survey on Deep Recurrent Neural Networks for Hyper Spectral Image Classification

G. Elayaroja and J.C. Miraclin Joyce Pamila
Department of Computer Science and Engineering,
Government College of Engineering, Coimbatore - 641 013, Tamil Nadu
E-mail : rojagovindan@gmail.com, miraclin@gct.ac.in

Abstract

Hyperspectral image processing has been a very dynamic area in remote sensing and other applications in recent years. Traditional methods, have shown promising results in hyperspectral image classification. Such methodologies, nevertheless, can lead to information loss in representing hyperspectral pixels, which intrinsically have a sequence-based data structure. A recurrent neural network (RNN), an essential branch of the deep learning class, is mightily propose to manage successive data. RNN framework has been proposed for hyperspectral image classification. Specifically, our RNN makes use of a recently designed activation function, parametric rectified tanh (PRetanh), for hyperspectral successive data analysis equivalent the familiar tanh or rectified linear unit. The proposed activation function makes it possible to use fairly high learning rates without the risk of divergence during the training procedure. Moreover, a modified gated recurrent unit, which uses PRetanh for hidden representation, is adopted to construct the recurrent layer in our network to efficiently process hyperspectral data and reduce the total number of parameters.

Keywords: *Convolutional neural network (CNN), Gatedrecurrent unit (GRU), Hyperspectral Imageclassification, Long short-term memory (LSTM), Recurrent neural network (RNN).*

1.INTRODUCTION

Remote sensing can be defined as collection and interpretation of information about an object, area or event without any physical contact with the object. Aircraft and satellites are the common platforms for remote sensing of earth and its natural resources (Goetz et al., 1985). Digital image processing is a process where input image is processed to get output also as an image or attributes of the image. Main aim of all image processing techniques is to recognize the image or object under consideration easier visually. It is used of extracting some useful information from original image. It is collects and processes information from across the electromagnetic spectrum. The goal of hyperspectral imaging is to obtain the spectrum for each pixel in the image of a scene, with the purpose of finding objects, identifying materials, or detecting processes.Image CalibrationImage calibrationalso referred to as camera resectioning,estimates the parameters of a lens and image sensor of an image or video camera.Image Segmentation Image segmentation is the separation of an picture into regions or categories, which address to separate objects or parts of objects. Every pixel in a picture is place one of a number of these categories.

Image Extraction Image processing, form extraction begin from an opening set of moderated data and builds deduce values (features) intended to be instructive and non-redundant, facilitating the succeeding learning and generalization measure.

Multivariate Data Analysis It is refers to any statistical technique used to analyze data that arises from more than one variable. This essentially pattern loyalty where each condition, product, or division involves more than a individual variable.

Classification Model A classification technique (or classifier) is a methodical approach to construction classification design from an input data set. Examples contain decision tree classifiers, regulation-based classifiers, neural networks, support vector machine etc..,

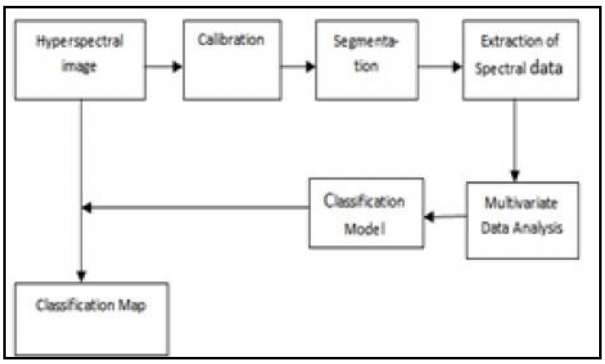


Fig.1 Steps Involved In Hyperspectral Image Classification

2. METHODS FOR HYPERSPECTRAL IMAGE CLASSIFICATION

2.1 Recurrent Neural Network (RNN)

An RNN is a class of artificial neural network that extends the conventional feedforward neural network with loops in connections. Unlike a feedforward neural network, an RNN is able to process the sequential inputs by having a recurrent hidden state whose activation at each step depends on that of the previous step. In this manner, the network can exhibit dynamic temporal behavior. The idea behind RNNs is to make use of sequential information. In a old-fashioned neural network we appropriate that all inputs (and outputs) are independent of each other. But for many work that's a very bade idea If you want to predict the next term in a significance you better know which words came before it. RNNs are called recurrent for they finish the same work for every constitute of a result, with the product being depended on the former computations.

$ht = \omega, (Wx_t + Uh_{t-1})$

2.2 Long Short-Term Memory Units (Lstms)

LSTMs help sustain the failure that can be back propagated through period and layers By continue a more constant failure, they assign recurrent nets to continue to study over many period steps (over 1000), thereby beginning a channel to link causes and result remotely. This is one of the central challenges to machine learning and AI, since algorithms are frequently confronted by environments where reward signals are sparse and delayed, such as life itself LSTMs hold information external the regular flow of the recurrent network in a gated cell. Information can be stored in, written to, or explain from a cell, much similar data in a computer's memory. The cell makes decisions about what to store and when to allow reads, writes and erasures, via gates that open and close. Unlike the digital warehousing on computers, however, these gates are analog, accomplish with element-wise augmentation by sigmoids, which are all in the rank of 0-1. Analoghas the benefit over digital of being differentiable, and therefore becoming for backpropagation.

2.3 Gated Recurrent nit (GRU)

A gated recurrent unit (GRU) is basically an LSTM without an output gate, which therefore fully writes the contents from its memory cell to the larger net at each time step. A GRU has two gates, a replace gate , and an update gate . Intuitively, the replace gate limit how to combine the modern input with the prior memory, and the update gate explain how much of the prior memory to keep around. If set the reset to all 1's and update gate to all 0's again arrive at our plain RNN model. The basic idea of using a gating mechanism to learn long-term dependencies is the same as in a LSTM.

2.4 Related works

FaridMelgani and Lorenzo Bruzzone [1] proposed a novel SVMs used to solve multiclass problems in hyperspectral data. This paper addresses the problem of the classification of hyperspectral remote sensing images by support vector machines (SVMs). First, a theoretical disputation and trial analysis scope at perception and assessing the potentialities of SVM classifiers in hyper dimensional feature spaces. Then, assess the effectiveness of SVMs with respect to conventional feature-reduction-based approaches and their performances in hyper subspaces of various dimensionalities. SVMs are a sufficient and powerful disjunctive to formal pattern recognition approaches (feature-reduction procedures combined with a classification process) for the classification of hyperspectral remote sensing data.

JiSoo Ham, Yangchi Chen, Melba M. Crawford and JoydeepGhosh [2] proposed a novel Random forest used to improved generalization of the classifier in analysis of hyperspectral data. This paper investigates two approaches based on the concept of random forests of classifiers implemented within a binary hierarchical multiclassifier system, with the goal of achieving improved generalization of the classifier in analysis of hyperspectral data, especially when the amount of training data is restricted. A modern classifier is proposed that incorporeal bagging of manege samples and adaptative random subspace form selection within a binary hierarchic classifier (BHC), such that the number of features that is choose at each node of the tree is dependent on the amount of combined training data.

Yushi Chen, Hanlu Jiang, Chunyang Li, XiupingJia, and PedramGhamisi [3] proposed a novel CNN used to extract effective spectral–spatial features of hyperspectral image. The proposed approximate employs several convolutional and pooling layers to citation deep form from HSIs, which are nonlinear, discriminant, and invariable. These form are useful for image classification and target discernment. More importantly, a 3-D CNN-

supported FE model with combined regularization to extract effective spectral–spatial form of hyperspectral images. In summary, to address the HSI FE and classification problem with restricted training example, an judgment of big network with valid constraints. The proposed model can be combined with post-classification processing to enhance mapping performance. It deserves to be investigated as a possible future work.

HaoboLyu and Hui Lu [4] proposed a novel for Improved LSTM model to learn the binary differencing information for multi-temporal remote sensing data. In this paper, a novel change detection method learned from Recurrent Neural Network with transferable ability is proposed. The proposed method, which is based on an improved Long Short Term Memory (LSTM) model, aims at: 1) learning a novel change detection rule to distinguish changed regions with high accuracy; 2) analyzing a new target data with transferable ability from learned change rule; 3) learning the differencing information and detecting the changes independently without any classifiers. In the process of learning the change rule, a core memory cell is utilized to detect and record the differencing information in multi-temporal images.

Adolfo Martinez-UsoFilibertoPla Jose Martinez Sotoca and Pedro Garcia-Sevilla [5] propose a novel for an approach is used to reduce data redundancy and nonusefulinformation among image bands. The proposed method is based on a hierarchical clustering structure to group bands to minimize the intracluster variance and maximize the intercluster variance. This aim is pursued using information measures, such as distances based on mutual information or Kullback–Leiblerdivergence, in order to reduce data redundancy and nonuseful information among image bands. Joseph C. Harsanyi and Chein-I Chang [6] propose a novel for an approach for simultaneously reducing hyperspectral data measurement and arrange the hyperspectral image. Most applications of hyperspectral imagery need processing techniques which realize two fundamental goals: 1) detect and classify the constituent materials for each pixel in the scene; 2) reduce the data volume dimensionality, without loss of critical information, so that it can be processed efficiently and assimilated by a human analyst.

Pal, M. and Foody, G. M. [7] proposed a novel for Feature selection is valuable analysis operations for classification by a SVM. SVM are attractive for the classification of remotely sensed data with some claims that the method is insensitive to the dimensionality of the data and so not requiring a dimensionality reduction analysis in pre-processing.Here, a sequence of classification analyses with two hyperspectral sensor data sets reveal that the correctness of a classification by a SVM does modify as a function of the many of features used. Critically, it is shown that the correctness of a classification may decline significantly (at 0.05 level of statistical significance) with the augmentation of features, especially if a small training example is used.This foreground a dependence of the correctness of classification by a SVM on the measurement of the data and so the potentially value of undertaking a feature selection analysis previous to classification.

Yushi Chen, Member, Zhouhan Lin, Xing Zhao, Student Member, Gang Wang, and YanfengGu, [8] propose a novel for Deep learning framework to merge the spectral and spatial information approach for highest classification accuracy. Classification is one of the most familiar topics in hyperspectral remote sensing. In the last two decades, a vast numeral of methods were intend to share with the hyperspectral data classification problem. However, most of them do not hierarchically extract deep features. In this paper, the conception of deep learning is insert into hyperspectral data classification for the first period. A novel deep learning framework to merge the two features, from which get the highest classification accuracy. The framework is a crossbreed of principle component analysis (PCA), deep learning structure, and logical regression. Specifically, as a deep learning architecture, stacked auto encoders are aimed to get useful high-level features.

SujuRajan, Joydeep 3 Ghosh, and Melba M. Crawford [9] propose a novel for an active learning access for effectively updating classifiers build from trivial quantities of labeled data. an active learning technique that efficiently updates existing classifiers by using fewer labeled data points than semi supervised methods. Further, unlike semi supervised methods, our proposed technique is well suited for learning or adapting classifiers when there is substantial change in the spectral signatures between labeled and unlabeled data. Thus, our active learning approach is also beneficial for distribute a series of spatially/secularly related images, wherein the spectral signatures modify across the images Our interfoliate semi supervised active learning process was standard on both divide and spatially/temporally related hyperspectral data sets.

Table 1 Summarization of Literature Survey

Sl. NO	TITLE	METHODS	ADVANTAGE	DISADVANTAGE
1	**Classification of Hyperspectral Remote Sensing Images with Support Vector Machines**	Feature-reduction Pattern recognition	More classification accuracy, Computational time, Stability to parameter setting.	SVM gives some information loss in hyperspectral data.
2	**Investigation of the Random Forest Framework for Classification of Hyperspectral Data.**	Binary hierarchical classifier (BHC) classification and regression trees (CART)	Improving generalization Greater Diversity higher accuracies	Random forest gives some data loss in hyperspectral data.
3	**Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks**	Feature extraction (FE)	CNN can extract the spectral–spatial features effectively, High classification accuracy	Limited training samples, Problem of overfitting.
4	**Learning A Transferable Change Detection Method by Recurrent Neural Network**	Multi-temporal images.	Higher accuracy for change detection, Effective transferable ability.	Less training time.
5	**Clustering-Based Hyperspectral Band Selection Using Information Measures**	Hierarchical clustering Kullback–Leibler divergence	Performance increases in image classification tasks. Higher accuracy	Band dissimilarity space plays in the problem of classification.
6	**Hyperspectral Image Classification and Dimensionality Reduction: An Orthogonal Subspace Projection Approach**	Minimum distance and maximum likelihood classifiers Principal components analysis(PCA)	To solve a particular detection and classification problem. Higher accuracy	Suffer from the mixed pixel problem.
7	**Feature selection for classification of hyperspectral data by SVM**	Feature selection	High potential rate. Higher classification accuracy. Computationally efficient	One problem often noted in the classification of hyperspectral data is the Hughes effect or phenomenon

8	**Deep Learning-Based Classification of Hyperspectral Data**	Stacked autoencoder(SAE) Logistic regression Support vector machine (SVM).	Highest accuracy when compared with other feature extraction methods. Both SAE-LR and SVM have proved the effective result	Testing time efficiency is less. Less training time.
9	**An Active Learning Approach to Hyperspectral Data Classification**	Semi supervised learning Active learning	Using very few labeled data points. Provide a high performance. Better learning rates.	Real life problems have large amount of unlabeled data.
10	**Semi-Supervised Neural Networks for Efficient Hyperspectral Image Classification**	Transductive SVM, and Laplacian SVM	Higher accuracy, Higher performance	computationally more expensive

3. CONCLUSION

In this survey, a novel RNN model for hyperspectral image classification is proposed by the observation that hyperspectral pixels can be regarded as sequential data. Specifically, a newly designed activation function PRetanh for hyperspectral data processing in RNN, providing an opportunity to use fairly high learning rates without the risk of getting stuck in the divergence. Furthermore, a modified GRU with PRetanh was developed to effectively analyze hyperspectral data. For hyperspectral image classification, recurrent network was shown to provide statistically higher accuracy than SVM-RBF and CNN. It considers the intrinsic sequential data structure of a hyperspectral pixel for the first time, representing a novel methodology for better understanding, modeling, and processing of hyperspectral data. In the future, further experiments will be conducted to fully substantiate the features of deep RNN for hyperspectral image processing, providing more accurate analysis for remote sensing applications, such as transfer learning for remote sensing big data analysis and change detection.

REFERENCES

[1] LichaoMou, Student Member IEEE, PedramGhamisi, Member IEEE, AndXiao Xiang Zhu, Senior Member IEEE, "Deep Recurrent Neural Networks For Hyperspectral Image Classification"IEEE Transactions On GeoscienceAnd Remote Sensing, Vol. 55, No. 7, July 2017.

[2] F. Melgani And L. Bruzzone, "Classification OfHyperspectral Remote Sensing Images With SupportVector Machines", IEEE Transactions Geo science And Remote Sensing, Vol. 42, No. 8, Aug 2004, pp. 1778- 1790.

[3] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the Random Forest Framework For Classification Of Hyperspectral Data", IEEE Transaction Geoscience And Remote Sensing, Vol.43, No.3, Mar 2005, pp. 492-501.

[4] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks", IEEE Transaction Geoscience And Remote Sensing, Vol. 54, No. 10, Oct 2016, pp. 6232-6251.

[5] H. Lyn, H. Lu, and L. Mou, "Learning a Transferable Change Rule From a Recurrent Neural Network for Land Cover Change Detection", Remote Sensing, Vol. 8, No. 6, 2016, pp.506.

[6] Adolfo Martinez-Uso, FilibertoPla, Jose Martinez Sotoca, and Pedro Garcia-Servilla, "Clustering-Based Hyperspectral Band Selection Using Information Measures", IEEE Transactions on Geoscience And Remote Sensing, Vol. 45, No. 12, December 2007.

[7] C. Joseph Harsanyi, Member IEEE, and Chein-I Chang, Senior Member IEEE, " Hyperspectral Image Classification and Dimensionality Reduction:

An Orthogonal Subspace Projection Approach", IEEE Transactions on Geoscience And Remote Sensing, Vol. 32, No. 4, July 1994.

[8] M. Pal and G.M.Foody, "Feature Selection for Classification of Hyperspectral Data by SVM", IEEE Transactions on Geoscience And Remote Sensing, Vol. 48, 2010, pp. 2297-2307.

Vehicular Air Pollution Monitoring in Traffic Area Using PIC16F877A

V.S. Esther Pushoam and S. Kumaresan
Department of Computer Science and Engineering
Government College of Technology, Coimbatore - 641 013, Tamil Nadu
E-mail: sthersrj@gmail.com

Abstract

Air Pollution is one of the major risk factors in India, due to that 1.2 million death takes place every year. The increase in the number of vehicles leads to heavy road traffic leading the way to high emission rate. Major pollutant from the vehicles are Carbon Monoxide (CO), Oxides of Nitrogen (NOx) and Particulate Matter (PM). Existing approach of measuring vehicle emission is a tedious process because it involved with various expensive instruments or simulation model. Since it is not efficient way to measure vehicle emission in real time, automation system for air quality monitoring in traffic area was developed. National Ambient Air Quality Standard (NAAQS) was studied to compare the pollutant level to select the threshold value. In proposed IOT-based approach, sensor array will be used to measure the CO,NO2 ,O3 concentration level at traffic area, the data will be compared with threshold value based on Air Quality Index(AQI). The location can be stamped using GPS. If the sensor data exceeds, the alternative route can be suggested using Google Map will be implemented. The data has been send through server through wireless network , the Master- slave network topology adapted for communication of data to improve the performance of the system.

Keywords: *Air Quality Index, GPS, Google Map, GSM, Master – Slave Approach, IOT, NAAQS, Vehicle emission, Sensor array*

1. INTRODUCTION

Air Pollution is one of the major problems in India. Recently Delhi was most polluted city in India. One of the factor of air pollution due the increase in vehicles growth. The exhaust emission of motor vehicles are majorly CO ,PM,HC, NOx, These pollutant introduced various health effects like asthma, heart disease, lung diseases. Nearly 75% of people in India exposed to air pollution everyday. To reduce the emission rate from the vehicle government of India Adapted EURO standard which describe the limited range of pollutant that can emit from different types of vehicle. Air Quality has been monitored by Nation Ambient Monitoring Programme, under which several agent are working on including state and central government. The air quality monitoring station was located in particular location and sample of the air was collected from different zone like industrial, traffic , and residential area. The pollutant vale was measured from the sample[6]. This approach include with various expensive instruments .someexisting approaches that used to measure the air quality in traffic areas are The vehicular emission was measured with mathematical approach using formula, the emission factors can vary based on speed , distance , fuel type of the vehicle. The other approach was international vehicle emission IVE model to estimate the emission level. This model highly used to estimate the emission in India because of its features that support different driving modes, Meteorologicalvariables, fueltype[5].Another modelling approach that simulated the traffic model using VISSIM and emission model using VERSIT+[9].The next expensive model that uses athelometer for Black carbon, CO analyser for CO, these instruments are highly expensive[2].Mathematical approach also used, the emission factor was calculated based on different parameter like speed, fuel type, distance[1]. These approaches are not suitable for the real time monitoring, hence Internet of things approach was selected to monitor the environment and update the user to aware of pollution level in their route and act accordingly. In real time few sensor based approaches also used. The Mobile-DAQ unit gathers air pollutants levels (CO, NO2, and SO2), and packs them in a frame with the GPS physical location, time, and date. The frame is subsequently uploaded to the GPRS-Modem and transmitted to the Pollution-Server via the public mobile network.Air-Pollution-Index: Function to convert the raw pollutant level received from each Mobile-DAQ to pollution standards called air quality index (AQI) using the formula The pollution standard is defined according

the air quality standards of a particular region[10].The sensor Node has been deployed on the top of the vehicle in their experiment they have used Metal oxide Semiconductor(MOS) sensor to measure CO,NO2,Ozone and Electrochemical sensor for CO . The Electrochemical senor have more linearity response than MOS, but Electrochemical sensor are expensive than MOS[8].Participatory sensing (PS) combines the use of everyday mobile devices, such as cellular phones, GPS technology and location-based services, and sensors to form interactive, bidirectional mobile sensing information systems. The environmental data are collected by a set of individual external sensors integrated in a board and working like one single sensor device with multi-sensing features. These sensors, integrated through the cellphone via Bluetooth, the Bluetooth interface communicates with the acquisition board to receive the air quality measurements. A connection is opened with the Arduino Module through which the environmental data are constantly received. And the data can be transferred to server through UDP. Google Web Toolkit and Web 2.0 tools used for Visualization interface . The pollution level; are classified as low,medium ,high and they are differentiated with different colors and displayed on Graphical user interface[7].

2. SYSTEM ARCHITECTURE

The sensor node used to detect the pollution level in the environment . The node has CO metal oxide gas sensor, NO2 and O3 Electro chemical gas sensor. There are five types of gas sensor available they are electrochemical, infrared, semiconductor, catalytic bead and photo ionization.

Table 2.1 Summary of Gas Sensor Technology

Category	Electrochemical Sensor	Metal Oxide Sensor	Photo Acoustic Sensor	Conventional IR Sensor
Low end accuracy	Good	Good	Poor	Good
High end accuracy	Good	Good	Better	Better
Temperature sensitivity	Better	Poor	Best	Poor
Humidity sensitivity	Poor	Poor	Poor	Poor
Life time	Poor	Better	Better	Best
Calibration frequency	Poor	Better	Poor	Better

PIC16F877A Microcontroller, GSM module , Battery power were used for single node. Node sends information to the server , after the data processing, the information will be mapped into the map.

2.1 MQ-7 –CO Sensor

Sensitive material of the MQ7 gas sensor is Tin oxide. It has lower conductivity in clean air. It is used to detect the carbon Monoxide (CO)level , It has two circuit Heater circuit and output circuit. Heater circuit works on cycle of high and low temperature. It detects CO at low temperature and cleans other gasses absorb at high temperature,the sensors conductivity is higher when the concentration of gas rising.

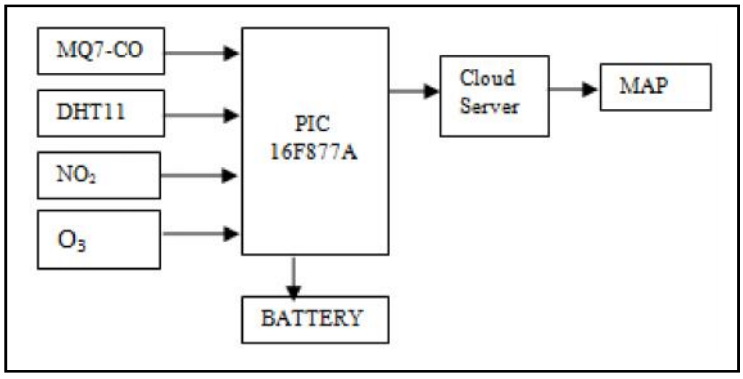


Fig.1 Node Architecture

Each node has CO, Temperature and Humidity, NO2, O3 sensor send data to cloud through PIC, from the cloud after the data processing , the data will be send to MAP.

2.2 DHT11 Sensor

It is a temperature and humidity sensor calibrated with digital signal output. The humidity measured through resistive type humidity measurement component and temperature measure through negative temperature coefficient (NTC) temperature measurement component. It has fast response, high reliability, long term stabilty.

2.3 NO2 Sensor

MICS 2710 is a NO2sensor detection range upto 5 ppm. When the sensor was heated through electricity, it detects the gases by showing variation in the resistance level. The features of this sensor are fast thermal response, wide detection range, low heater current, high sensitivity.

2.4 O3 Sensor

MICS 2610 is a O3sensor detection range upto 10000 ppb. It has a tin oxide as a sensitive material. It operated on the heating principle, the measurement has be taken based on the changes in the resistance level.

2.5 Micro Controller

40 pin Pic16F877A was used. It operates on 20MHZ frequency, It has 8 analog input channel. It has 8 10-bit ADC channel. It has Data Memory Upto 368 Bytes, EEPROM upto 256 Bytes, It supports serial communication.

In this architecture the senor array collects data from the environment and send that information to the cloud corresponding location on the map. The Location was marked as safe or not safe based on AQI Value based on the information user can change their Route. Master –Slave network topology was used in this system. This node can be fixed in every possible route, hence if every single node communicate with server there might be data congestion which leads to data loss or failure of the system, In this system , Master – Slave approach was used . In this approach , In a region one node acts as a master and other node in that region acts as a slave , slave node sends information to the master node, and the master node sends that information to cloud server using GSM.

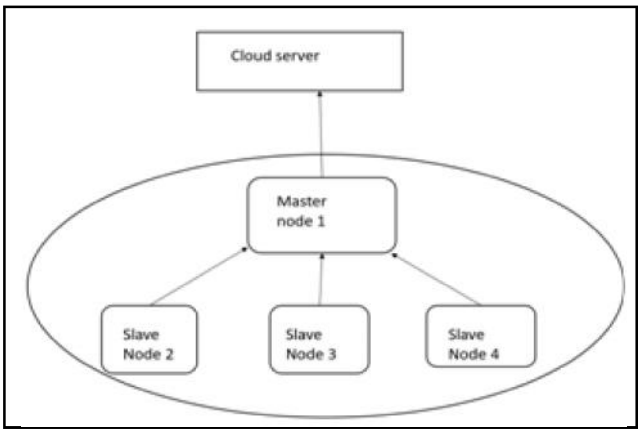


Fig.2 Network Topology

Master – Slave approach, one node acts as a master in a region other nodes are slave node in the same region. The slave node sends data to master . Master sends information to the server through GSM.

3. EXPERIMENT AND RESLUT

The sensor node implemented on all possible route, the node has to be deployed in the center of the node , possibly traffic signal . Each node sends its information to the server.The readings are checked with Air Quality Index value and categorized.

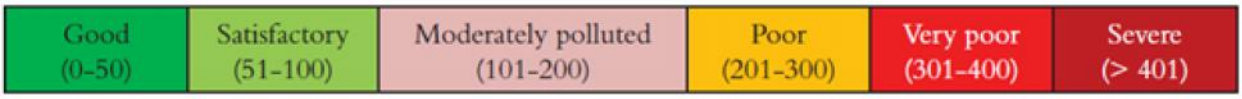


Fig.3 AQI Colour code – shows the pollution range

The colour code indicates the the quality of the air The AQI value calculated as

AIR QUALITY INDEX = POLLUTION LEVEL *100
POLLUTION STANDARD

The data from the node was send to the Cloud server , the comparison of pollution level at three routes was made , the pollution level of the route was mapped to Google map . The pollution level also indicated with color code.

Using this approach , the air pollution can be monitored in traffic area and pollution level status of that route is updated in real time on map. Hence user can avoid that route or take an alternative route to reach their destination safer. Since vehicles are spread across different route, there might be significant reduction in pollution level, The network topology used in this approach is a master slave approach which reduced data traffic to server, because only few node communicating directly with the server. However if all the master node access the server at same time there might be bottleneck, it can be avoided by providing timeslot for master node.

4. CONCLUSION.

This system will be helpful for the government to monitor the airpollution in the traffic area , and intimate the pollution status of that route to the user in real time, using this approach government can provide better environment to the people, however this approach have some limitation, the nodes are need to be deployed large number, energy efficiency of the node also need to be considered , it can be improved by using solar energy. The natural parameters like wind , Humidity can affect the accuracy of the sensor. In such a case For real time approach it is suitable to use electrochemical sensor since it has high accuracy.

REFFERNCE

[1] M Tiwari, SP Shukla, NK Shukla, RB Singh, N Mumtaz, VK Gupta and V Kumar., "Emission Profile of Pollutants Due to Traffic in Lucknow City, India", International Research Journal of Public and Environmental Health Vol.1. No.7, 2014, pp.150-157.

[2] D V Mahalakshmi, P Sujatha,P V Naidu and V M Chowdary, "Contribution of Vehicular Emission from Urban Air Quality : Result From Public Strike in Hyderabad", Indian Jornal of Radio & Space Physics,Vol.43, Dec 2014, pp-340-348.

[3] P. Partheeban, R. Rani Hemamalini and H. Prasad Raju, "Vehicular Emission Monitoring Using Internet GIS, GPS and Sensors", International Conference on Environment, Energy and Biotechnology,IPCBEE, Vol.33, 2014, .pp-81-85.

[4] WataruTsujita, Akihito Yoshino, Hiroshi Ishida, ToyosakaMoriizumi, "Gas Sensor Network for Air-Pollution Monitoring", ELSEVIER Sensors and Actuators B 110 (2005), 2015, pp.304-311.

[5] PramilaGoyal, Dhirendra Mishra and Anikender Kumar., 2013,"Vehicular Emission Inventory of Criteria Pollutants in Delhi", online journal: http://www.springerplus.com/content/2/1/216.

[6] Central Pollution Control Board Ministry Of Environment & Forests, "National Ambient Air Quality Status & Trends in India -2010".

[7] D. Mendez, A. J. Perez, M. A. Labrador and J. J. Marron, "P-sense:A Participatory Sensing System for Air Pollution Monitoring And Control,"in Proc. IEEE Int. Conf. Pervasive Comput. Commun.Workshops (PERCOM Workshops), pp.344-347.

[8] KeHu, VijaySivaraman, Member, IEEE, BalancaGallegoLuxan and AshfaqurRahman,senior Member, IEEE, "Design and Evaluation of Metropolitan Air Pollution Sensing System", IEEE Sensor Journal, Vol.16, No.5, March 1, 2016, pp-1448-1459.

[9] Mohamed Mahmod, Bart van Aerm, Rattaphol, Ronald delange, "Reducing Local Traffic Emission At Urban Intersection Using ITS Countermeasures", IET Intelligent Transport System., 2013, Vol.7, No.1, 2013, pp.78-86.

[10] A. R. Al-Ali, Member, IEEE, Imran Zualkernan, and FadiAloul, Senior Member, IEEE, "A Mobile GPRS-Sensors Array for Air Pollution Monitoring", IEEE SENSORS JOURNAL, Vol.10, No.10, pp-1666-1671.

[11] SivaShankarChandrasekaran, Sudharshan Muthukumar and Sabeshkumar Rajendran, "Automated Control System for Air pollution Detection in Vehicles", 4th International Conference on Intelligent Systems, Modelling and Simulation, 2013, pp-49-51.

[12] David Hasenfratz, Olga Saukha, ChristophWalser, ChristophHueglin, Martin Fierz, TabitaArna, Jan Beutel, Lothar Thiele, "Deriving High-resolution Urban Air Pollution Maps using Mobile Sensor Nodes", Elsevier B.V, 2014, pp-1574-1192.

[13] Kgoputjo Simon Elvis Phala, Anuj Kumar, and Gerhard P. Hancke, Senior Member, IEEE, "Air Quality Monitoring System Based on ISO/IEC/IEEE 21451 Standards", IEEE Sensors Journal, Vol.16, No.12, June15, 2016, pp-5037-5045.

[14] O. Pummakarnchanaa,N. Tripathia, J. Dutta, "Air Pollution Monitoring and GIS Modeling: A New Use of Nanotechnology Based Solid State Gas Sensors", Elsevier Science and Technology of Advanced Materials, Vol.6, 2005, pp.251-255.

Survey on Finding Related Forum Post

A.K. Ajithkumar and J.C.Miraclin Joyce Pamila
Department of Computer Science and Engineering,
Government College Of Technology, Coimbatore - 641 013, Tamil Nadu

Abstract

This article provides a comprehensive overview of finding related forum posts. Forums in online communities contribute rich content, concerning diverse problems by utilizing other users experience. However, browsing a very large count of forum posts is frustrating and time consuming. Finding related forum posts is not trivial, because the content of the forum posts are written in natural language and it's usage in forum posts does not follow any particular structure. The web forums follow different styles and architecture according to the context of web site. This makes, finding related forum posts more difficult. This article explains some of the existing techniques used for finding related forum posts in a comprehensive manner. These methods are classified according to the specific techniques used. It also contains some of the directions of the field of study and we classify this direction of study according to two dimensions: finding related forum posts through statistical or lexical methods and finding related forum posts through semantic similarity. We believe that this article will be use full for the researchers who are interested in the field of information retrieval and web forums.

1. INTRODUCTION

On the era of Web 2.0, the world wide web is shifting from static to dynamic, where the empha-size is on user generated content, usability and interoperability for end users. Most of the Web 2.0 websites offer internet forums which is used to share information from all around the world. With the flowering of Web 2.0, social collaborative applications such as Wikipedia, YouTube, Facebook etc. begin to flourish, and there have been an increasing number of Web information services that bring together a network of people to answer questions posted by other folks.

An Internet forum, is an online discussion site where people can hold conversations in the form of posted messages. Also, depending on the access level of a user or the forum set-up, a posted message might need to be approved by a moderator before it becomes visible. Forums have a specific set of jargon associated with them. for example, a single conversation is called a "thread", or topic. A discussion forum is hierarchical or tree-like in structure. A forum can contain a number of subforums, each of which may have several topics. Within a forum's topic, each new discussion started is called a thread, and can be replied to by as many people as so wish. Depending on the forum's settings, users can be anonymous or have to register with the forum and then subsequently login in order to post messages. On most forums, users do not have to login to read existing messages.

People, visit forum sites from every corner of the world, with different intentions. Most of the time,they discuss specific topics related to politics, health, sports, movies and music. Some forum sites are interested in sharing information. In those type of websites, a specific academic or research topic will be elaborated in depth and people will discuss issues, regarding these specific fields. Stackoverflow for programmers is an excellent example for this kind of forum. Another kind of forums are interested in clearing doubts, in which users directly ask questions in the community.This is referred to as the community-based question answering services (cQA). In these communities, anyone can ask and answer questions on any topic, and people seeking information are connected to those who know the answer. As answers are usually explicitly provided by human and are of high quality, they can be helpful in answering real world questions.

While the motivation for users to participate in discussion boards varies, in many cases, people would like to use discussion boards as problem-solving platforms. Users post ques-tions, usually related to some specific problem, and rely on others to provide potential answers. Numerous commercial organizations such as Dell and IBM directly use discussion boards as problem-solving solutions for answering questions and discussing needs posed by customers.

1.1 Motivation

The experience of people are used in order to seek solutions in web forums. Even businesses have started to use web forums to connect and support their customers. The forums range from domains like health (e.g., Medhelp), law (e.g.,ExpertLaw) and technology (e.g.,HP support forum) are available in internet. The organization of the forum post categorization is based on topic classes is done by most of the web forums. However, since browsing a very large number of posts is frustrating and time-consuming, most forum sites offer keyword search capabilities. Yet, keyword search may not result in a complete set of related posts since the selection of the right keywords is difficult for common users. A functionality that can be added to better support users, is to automatically provide them with a collection of related documents if they have already identified a forum post of interest. This will help them to find exact content they are looking for, without formulating complicated queries or perform complicated, long browsing. There are multiple studies that have been done on the field of finding related forum posts. And most of these studies are based on lexical methods. In next section, we provide classification of techniques that have been used to address issue of finding related forum posts.

2. TOWARDS FINDING RELATED FORUM POSTS

Since, most of the forums are cQA based, there exists large area of work for finding related ques-tions or questions-answer pairs. The classification of these techniques can be done in following way.

Based on thread similarity

Based on syntactic structure of questions

Based on thread post reply structure

Based on different combination of NLP features These techniques are elaborated in following sections.

2.1 Based on Thread Similarity

These methods exploits the specific structural features of web forums in order to find similar forum posts.

4 Uses question-answer pair relationships to find the similar questions. Their proposed method assumes “if two questions are very similar, then the questions connected to the answers should semantically similar, even though the two questions may be lexically very different. The language modelling methods [2] worked better for their proposal. Also they have introduced their own similarity measure technique known as LM-SCORE, and used to cluster question answer pairs. Here author believes that, current knowledge-bases are not good enough to find semantic similarity and using manual rules are too expensive. So they followed statistical and NLP(Natural Language Processing) techniques which is believed to be work well as long as sufficient data is available. They have utilized the word translational probabilities of the text to find semantic similarity. They have utilized the large collection of Q&A archives to infer these results from the study.

In [3], the authors address the problem of suggesting questions to the users which is sim-ilar to already present ones. They have identified bag of words [4] wont perform better since it does not consider semantics while comparing with related questions. They have proposed Topic enhanced Translation-based Language Model (TopicTRLM) which fuses both the lexical and latent semantic knowledge. But they have not tried to measure quality of the suggested questions, and missed the opportunity to improve the system with feedback. The topic models works by assigning set of latent topic distributions to each words. Their proposed TopicTRLM model fuses the latent topic information with lexical information and for this they have used the famous Latent Dirichlet Allocation(LDA).

2.2 Based on Syntactic Structure of Questions

Here, syntax of sentences in the post is considered for finding related content. As syntax and semantics are closely related this might lead to similar forum posts.

ω proposed a new retrieval framework based on syntactic tree structure to tackle the similar question matching problem.

They were interested in finding related questions from the web forums. Since people on the internet are from different part of the world, not only their language but also the cultural differ-ences affect their vocabulary and they make same questions in an entirely different way. This phenomenon caused by both language barrier and cultural differences have made the problem of finding similar questions non-trivial.

For example, "how can I lose weight in a few month?" and "are there any ways of losing pound in a short period?" are two related questions asking for ways of losing weight, but they have no common words or structure. In this way finding similar questions task, become so difficult. Also BOW based methods won't work well in these situations.

The tree kernel function[6] has been effectively applied in some areas like question classi-fication, and it is stretched to their technique.

The tree kernel metric is used to find the distance between two sentences, but there is some issues with this method, the first issue is that tree kernel function just depends on number of common sub-trees between the sentences, which may be not an indicator of similarity and the next one is two evaluated sub-trees have to be identical to allow further parent matching, for which semantic depictions cannot acceptable in well. To remedy the second issue, the Shallow Semantic Tree Kernel (SSTK) was proposed in , where Predicate Argument Structures (PAS) are exploited to take dependencies into account. However, it was noted to be computational expensive for real world applications. They further enhanced the system performance by incorporating semantic features to the lexemes and the answer matching module in the framework. Unlike other sys-tems, their model does ¬¬not rely on training, making it easily portable to other similar retrieval systems.

In [7]'s work, they addressed the problem of segmenting multi-sentence questions, which are complemented with various contexts, thus making difficulties in finding related questions. They have presented a new segmentation approach using graph based model. Even though their segmentation techniques significantly improves question matching performance, they have not used features of answer pairs for particular questions for the segmentation of questions.

2.3 Based on Thread Post Reply Structure

Most forums have specific structures. We can exploit thread structure in order to find related forum posts. Since most of the forums follow thread structure we can apply similarity measures based on threads. The study based on threads has been done by [8] and they address the problem of finding similar threads to a given thread.

Towards this, they propose a novel methodology to evaluate similarity between discussion threads. They segmented forum to set of weighted overlapping components using thread struc-ture property.

Their method considered a thread as a single large document and by doing that they exploits techniques that will work based on large document mode. The large document is formed by putting all the text from component posts. Now, the similarities between threads estimated as the similarities between such documents as assessed using one of the popular text similarity measures that use the bag of words (BOW) model. They experimented with measures such as cosine similarity of term frequency vectors, cosine similarity of tfidf vectors and Jaccard similarity co-efficient. Similarity measures used by them is concerned with how well threads content contained mutually within each other document.

They have calculated pair wise thread similarity, by modelling threads as a set of weighted overlapping components. The quantification was by lexical similarity between components from the threads under consideration.

Specifically, they use the post-reply component that has been found to be beneficial in dis-cussion forum search, in addition to using separate posts as components; they proved that their similarity computation can run in polynomial time for this choice of component types. They de-fined numerous intuitive methods that use thread retrieval techniques and direct thread compar-ison and evaluated their approaches against them. Through an extensive series of experiments on real world data, they established the effectiveness of their technique on popular similarity measures such as NDCG, MAP, Precision and MRR. Specifically, their techniques are seen to outperform the baselines by approximately 10each of the measures. Even though they did not considered the semantical features the results were promising. But they have missed out some features of web forums like authorships, time, etc. Also more sophisticated structural compo-nents have not been taken in to account which could potentially improve retrieval quality.

2.4 Based on Different Combination of NLP Features

The NLP features like number of question marks, 5w1h words[9],authorship, number and dis-tribution of

stop words can be effectively used in the domain of study. N-grams[10] is another famous NLP technique that is used in computational linguistics, which can be used in our con-text also.

3. STATISTICAL METHODS

Although the number of papers on finding related forum posts is small, related works are numer-ous. finding related forum posts is a specialization of finding related documents or information retrieval.

3.1 Classifying Questions and Answers

+ consider the problems of identifying question-related threads and their potential answers as classification tasks. In their paper, they considered the problem of finding questions and corresponding answers from web forums, And it is done in two parts. Their first task is to identify questions related to first post of web forums, and second task is to finding potential answers to that questions identified in first task.

For Question detection task, they describe and use several features other researchers have used previously (e.g., question mark, 5W1H words) as well as features that are borrowed from other fields (e.g., Ngram). For Answer Detection, they have used the features such as position of the answer post, authorship, N-gram, Query Likelihood Model Score (Language Model), etc.

They show that the use of N-grams and the combination of several non-content features can improve the performance of detecting question-related threads in discussion boards.Also They show that the number of posts a user starts and the number of replies produced and their positions are two crucial factors in determining potential answers.

3.2 Segmentation of Posts using Grammatical Features

In contrast to traditional approaches for finding related documents that perform content com-parisons across the content of the posts as a whole,[12] consider each post as a set of segments, each written with a different goal in mind. They advocate that the relatedness between two posts should be based on the similarity of their respective segments that are intended for the same goal, i.e., are conveying the same intention. This means that it is possible for the same terms to weigh differently in the relatedness score depending on the intention of the segment in which they are found. They have developed a segmentation method that by monitoring a number of text features can identify the parts of a post where significant jumps occur indicating a point where a segmentation should take place. The generated segments of all the posts are clustered to form intention clusters and then similarities across the posts are calculated through similarities across segments with the same intention.

Their method identifies and exploits post segments that convey similar author intentions. They presented several experiments regarding the right segmentation criteria, the effectiveness of the segmentation algorithms and the formation of intention clusters that prove that a rather intuitive concept, that of the author intentions to communicate a certain message, can be effec-tively captured by an automated process.

3.3 Statistical Similarity Measures

These are some statistical similarity measures that is used in information retrieval systems.

Cosine similarity with tfidf :[14]

Language modelling techniques:[15]

NDCG:[16]

MAP:[17]

MRR: [17]

4. SEMANTICAL METHODS

To improve the quality of content retrieved,the semantic has to be considered. It is possible only by conceptualizing the text. These are some works done trying to simulate the understanding power of algorithms.

4.1 Using External Semantic Network

In their work, [18] use lexical semantic knowledge provided by a well-known semantic network for short text understanding. Their knowledge-intensive approach disrupts traditional methods for tasks such as text segmentation, part-of-speech tagging, and concept labeling, in the sense that they focus on semantics in all these tasks. In their work, they followed three steps

namely text segmentation, type detection and concept labelling. Their techniques shows that external knowledge are important for conceptualizing text.

4.2 Using a Probabilistic Knowledge-base

Most text mining tasks, including clustering and topic detection, are based on statistical meth-ods that treat text as bags of words. Semantics in the text is largely ignored in the mining process, and mining results often have low interpretability. One particular challenge faced by such approaches lies in short text understanding, as short texts lack enough content from which statistical conclusions can be drawn easily.

In their research, [19] have used a probablistic knowledge-base to understand twitter data. They have found out probase entity from twitter feed and understand the concepts with Bayesian inference.For performance they build a trie to index the terms. Results shows that their approach is highly effective compared to traditional bag-of-words based statistical methods.

5. CONCLUSION

In this survey, we have covered techniques, which uses lexical methods for finding related forum posts. Although, these methods are giving promising results, they are strictly based on statistical similarity, and thus not giving relatedness based on meaning, because these techniques not trying to understand text. Most of them, are just utilizing lexical signals, like number of words or synonyms, in order to find related content.

If we need semantically related forums, we must understand the text and compare the con-cepts, rather than comparing statistical content features. For this purpose, we need to add se-mantic analyzer part to the system's framework which will do the understanding part. And here, understanding the text refers to "conceptualizing the terms and finding intention of the author". We could understand the text by using semantic networks like knowledge bases. Also Named Entity Recognition and Named Entity Disambiguation has to be done, to find exact intention of the author of the forum post.

REFERENCES

[1] J. Jeon, W. B. Croft and J. H. Lee, "Finding Similar Questions in Large Question and Answer Archives", in Proceedings of the 14th ACM International Conference on Information and Knowledge Management, ser. CIKM '05. New York, NY, USA: ACM, [Online]. Available: http://doi.acm.org/10.1145/1099554.1099572, 2005, pp. 84-90.

[2] J. M. Ponte and W. B. Croft, "A Language Modeling Approach to Information Retrieval", in Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '98. New York, NY, USA: ACM, [Online]. Available: http://doi.acm.org/10.1145/290941.291008, 1998, pp. 275–281.

[3] T. C. Zhou, C.-Y. Lin, I. King, M. R. Lyu, Y.-I. Song, and Y. Cao, "Learning to Suggest Questions in Online Forums", 2011.

[4] Y. Zhang, R. Jin and Z.-H. Zhou, "Understanding Bag-Of-Words Model: A Statistical Frame-Work", International Journal of Machine Learning and Cybernetics, Vol.1, No.1-4, 2010, pp.43-52.

[5] K. Wang, Z. Ming and T.-S. Chua, "A Syntactic Tree Matching Approach to Finding Similar Questions In Community-Based Qa Services", in Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '09. New York, NY, USA: ACM, [Online]. Available: http://doi.acm.org/10.1145/1571941.1571975, 2009, pp.187-194.

[6] D. Zhang and W. S. Lee, "Question Classification Using Support Vector Machines", in Pro-ceedings of the 26th Annual international ACM SIGIR conference on Research and devel-opment in informaion retrieval. ACM, 2003, pp.26-32.

[7] K. Wang, Z.-Y. Ming, X. Hu, and T.-S. Chua, "Segmentation of Multi-Sentence Questions: Towards Effective Question Retrieval in CQA Services", in Proceedings of the 33rd inter-national ACM SIGIR conference on Research and development in information retrieval. ACM, 2010, pp. 387-394.

[8] A. Singh, D. P, and D. Raghu, "Retrieving Similar Discussion Forum Threads: A Structure Based Approach", in Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR

12. New York, NY, USA: ACM, [Online]. Available: http://doi.acm.org/10.1145/2348283.2348305, 2012, pp. 135-144.

[9] T. Ikeda, A. Okumura, and K. Muraki, "Information Classification And Navigation Based On 5w1h Of the Target Information", in Proceedings of the 36th Annual Meeting of the Associ-ation for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 1998, pp. 571-577.

[10] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra and J. C. Lai, "Class-based n-gram Models of Natural Language", Computational Linguistics, Vol.18, No.4, 1992, pp. 467-479.

[11] L. Hong and B. D. Davison, "Swering in Discussion Boards", SIGIR Conference on Research and Development in Information Retrieval, ser. SI-GIR '09. New York, NY, USA: ACM, [Online]. Available: http://doi.acm.org/10.1145/1571941.1571973, 2009, pp.171-178.

[12] D. Papadimitriou, G. Koutrika, Y. Velegrakis, and J. Mylopoulos, "Finding Related Forum Posts Through Content Similarity Over Intention-Based Segmentation", IEEE Transactions on Knowledge and Data Engineering, Vol. 29, No. 9, Sept 2017, pp.1860-1873.

[13] A. Shtok, G. Dror, Y. Maarek, and I. Szpektor, "Learning from the Past: Answering New Questions with Past Answers", in Proceedings of the 21st International Conference on World Wide Web, ser. WWW '12. New York, NY, USA: ACM, [Online]. Available: http://doi.acm.org/10.1145/2187836.2187939, 2012, pp.759-768.

[14] S. Tata and J. M. Patel, "Estimating the Selectivity of TF-IDF Based Cosine Similarity Predi-Cates", ACM Sigmod Record, Vol.36, No.2, 2007, pp.7-12.

[15] R. Kneser and H. Ney, "Improved Clustering Techniques for Class-Based Statistical Language Modelling", in Eurospeech, Vol.93, 1993, pp.973-76.

Data Mining Techniques and its Application

B. Rajdeepa and D. Pavithra
Department of Computer Science,
PSG College of Arts and Science, Coimbatore - 641 014, Tamil Nadu
Email : rajdeepab@gmail.com, pavikoki555@gmail.com

Abstract

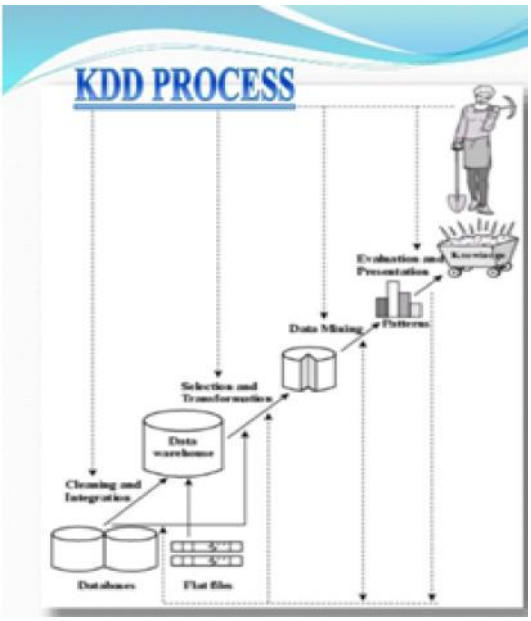
Data mining, also called knowledge discovery in databases, in computer science, the process of discovering interesting and useful patterns and relationships in large volumes of data. Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data. The practice of examining large pre-existing databases in order to generate new information. In data mining and then gradually moves on to cover topics such as knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis, and how to mine the Web. This paper also discusses some of the data mining applications and techniques.

Keywords: *Classification and prediction, Cluster analysis, Decision tree induction, Data mining applications and techniques, Knowledge discovery*

1. INTRODUCTION

The knowledge discovery in databases (KDD) process is commonly defined with the stages:

- Selection
- Pre-processing
- Transformation
- Data mining
- Interpretation/evaluation.[5]



Data Selection: Here, data relevant to the analysis are retrieved from various resources.

Pre-processing[edit]
Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

Data Transformation: In this step, data is converted or consolidated into required forms for mining by performing different operations such as smoothing, normalization or aggregation.

Data mining

Data mining, also called knowledge discovery in databases, in computer science, the process of discovering interesting and useful patterns and relationships in large volumes of data.

Pattern Evaluation: At this step, Attractive patterns representing knowledge are identified based on given measures.

Knowledge Representation: Present the mined knowledge to the user.

2. DATA MINING TECHNIQUES

2.1 Classification

This analysis is used to retrieve important and relevant information about data, and metadata. It is used to classify different data in different classes. Classification is similar to clustering in a way that it also segments data records into different segments called classes. A classic example of classification analysis would be our Outlook email.

2.2 Association Rule Learning

It refers to the method that can help you identify some interesting relations (dependency modeling) between different variables in large databases. This technique can help you unpack some hidden patterns in the data that can be used to identify variables within the data and the concurrence of different variables that appear very frequently in the dataset. Association rules are useful for examining and forecasting customer behavior. It is highly recommended in the retail industry analysis. This technique is used to determine shopping basket data analysis, product clustering, catalog design and store layout.

2.3 Clustering Analysis

The cluster is actually a collection of data objects; those objects are similar within the same cluster. That means the objects are similar to one another within the same group and they are rather different or they are dissimilar or unrelated to the objects in other groups or in other clusters. Clustering analysis is the process of discovering groups and clusters in the data in such a way that the degree of association between two objects is highest if they belong to the same group and lowest otherwise.A result of this analysis can be used to create customer profiling.

2.4 Regression Analysis

In statistical terms, a regression analysis is the process of identifying and analyzing the relationship among variables. It can help you understand the characteristic value of the dependent variable changes, if any one of the independent variables is varied. This means one variable is dependent on another, but it is not vice versa.It is generally used for prediction and forecasting.

All of these techniques can help analyze different data from different perspectives.

3. ADVANTAGES OF DATA MINING

3.1 Marketing / Retail

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have an appropriate approach to selling profitable products to targeted customers. Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

3.2 Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank, and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

3.3 Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipment and determine optimal control parameters. For example, semiconductor manufacturers have a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are a lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of the golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

3.4 Governments

Data mining helps government agency by digging and analyzing records of the financial transaction to build patterns that can detect money laundering or criminal activities.

4. DISADVANTAGES OF DATA MINING

4.1 Privacy Issues

The concerns about the personal privacy have been increasing enormously recently especially when the internet is booming with social networks, e-commerce, forums, blogs…. Because of privacy issues, people are afraid of their personal information is collected and used in an unethical way that potentially causing them a lot of troubles. Businesses collect information about their customers in many ways for understanding their purchasing behaviors trends. However businesses don't last forever, some days they may be acquired by other or gone. At this time, the personal information they own probably is sold to other or leak.

4.2 Security Issues

Security is a big issue. Businesses own information about their employees and customers including social security number, birthday, payroll and etc. However how properly this information is taken care is still in questions. There have been a lot of cases that hackers accessed and stole big data of customers from the big corporation such as Ford Motor Credit Company, Sony… with so much personal and financial information available, the credit card stolen and identity theft become a big problem. Misuse of information/inaccurate information is collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people. In addition, data mining technique is not perfectly accurate. Therefore, if inaccurate information is used for decision-making, it will cause serious consequence.

5. USEFUL APPLICATIONS OF DATA MINING

Data Mining is primarily used today by companies with a strong consumer focus-retail, financial, communication, and marketing organizations, to "drill down" into their transactional data and determine pricing, customer preferences and product positioning, impact on sales, customer satisfaction and corporate profits. With data mining, a retailer can use point-of- sale records of customer purchases to develop products and promotions to appeal to specific customer segments. Here is the list of 14 other important areas where data mining is widely used: Future Healthcare Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

5.1 Market Basket Analysis

Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behaviour of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

5.2 Education

There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

5.3 Fraud Detection

Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

5.4 Criminal Investigation

Criminology is a process that aims to identify crime characteristics. Actually crime analysis includes exploring

and detecting crimes and their relationships with criminals. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Text based crime reports can be converted into word processing files. These information can be used to perform crime matching process.

5.5 Bio Informatics

Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience. Applications of data mining to bioinformatics include gene finding, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

REFERENCES

[1] N.Verma, "Improved Web Mining for E-commerce Website Restructuring", In IEEE International Conference Computational Intelligence &Communication Technology, UP, India, Feb 2015, pp.155-160.

[2] N.Verma, D.Malhotra, M.Malhotra and J.Singh, "E-commerce Website Ranking Using Semantic Web Mining And Neural Computing" In Proceedings of International Conference on Advanced Computing Technologies and Applications, Mumbai, India, 2015. Procedia Computer Science, Science Direct. Elsevier, Vol. 45, pp.42-51.

[3] D.Malhotra, N.Verma, "An Ingenious Pattern Matching Approach to Ameliorate Web Page Rank", In International Journal of Computer Applications, New York, USA, FCS, Vol. 65, No 24, 2013, pp.33-39.

[4] D.Malhotra, "Intelligent Web Mining to Ameliorate Web Page Rank using Back Propagation Neural Network", In Proceedings of 5th International Conference, Confluence: The Next generation Information Technology Summit, Noida, India, IEEE, 2014, pp.77-81.

Analysis of Public-Key Cryptography for Wireless Sensor Networks Security

M. Infant Angel and R. Sudha
Department of Computer Science,
PSG College of Arts & Science,Coimbatore - 641 014, Tamil Nadu

Abstract

With the across the board development of utilizations of Wireless Sensor Networks (WSNs), the requirement for dependable security components these systems has expanded complex. Numerous security arrangements have been proposed in the area of WSN up until now. These arrangements are generally in view of surely understood cryptographic algorithmsWireless sensor systems comprise of self-ruling sensor hubs connected to at least one base stations.As Wireless sensor systems proceeds to grow,they end up plainly powerless against assaults and consequently the requirement for successful security mechanisms.Identification of appropriate cryptography for remote sensor systems is a critical test because of confinement of energy,computation capacity and capacity assets of the sensor nodes.Symmetric based cryptographic plans donot scale well when the quantity of sensor hubs increases.Hence open key based plans are broadly used. We show here two open - key based calculations. RSA and Elliptic Curve Cryptography (ECC) and discovered that ECC have a huge preferred standpoint over RSA as it diminishes the calculation time and furthermore the measure of information transmitted and put away.

Keywords: *Key Management, Public Key Cryptography, Security, Wireless Sensor Networks*

1.WIRELESS SENSOR NETWORK

Sensor networks refer to a heterogeneous system combining tiny sensors and actuators with generalpurpose computing elements. These networks will consist of hundreds or thousands of self-organizing, lowpower, low-cost wireless nodes deployed to monitor and affect the environment [1]. Sensor networks are typically characterized by limited power supplies, low bandwidth, small memory sizes and limited energy. This leads to a very demanding environment to provide security.

2. SECURITY REQUIREMENTS IN WIRELESS SENSOR NETWORK

The goal of security services in WSNs is to protect the information and resources from attacks and misbehaviour. The security requirements in WSN include: Confidentiality:

Confidentiality is hiding the information from unauthorized access. In many applications, nodes communicate highly sensitive data. A sensor network should not leak sensor reading to neighbouring networks.Simple method to keep sensitive data secret is to encrypt the data with a secretkey that only the intended receivers'possess, hence achieving confidentiality.As public key cryptography is too expensive to be used in the resource constrained sensor networks,most of the proposed protocols use symmetric key encryption methods.For symmetric key approach the key distribution mechanism should be extremely robust.

2.1 Authentication

Authentication ensures the reliability of the message by identifying its origin.In a WSN,the issue of authentication should address the following requirements:[1] communicating node is the one that it claims to be(ii)the receiver should verify that the received packets have undeniably come from the actual sensor node.For Authentication to be achieved the two parties should share a secret key to compute message authentication code(MAC) of all communicated data.The receiver will verify the authentication of the received message by using the MAC key.

2.2 Integrity

Integrityis preventing the information from unauthorized modification.Data authentication can provide data integrity also.

2.3 Availability

Availability ensures that services and information can be accessed at the time they are required. In sensor networks there are many risks that could result in loss

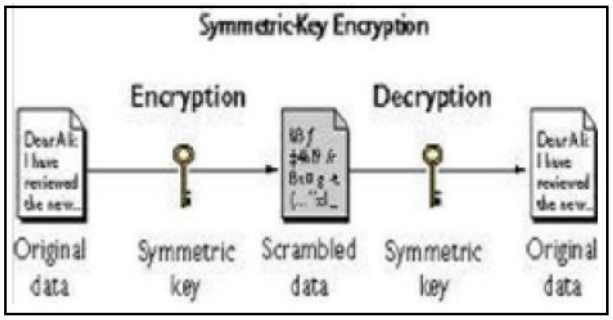
of availability such as sensor node capturing and denial of service attacks.

3. CRYPTOGRAPHY

Cryptography schemes are often utilized to meet the basic security requirements of confidentiality and integrity in networks. But as the sensor nodes are limited in their computational and memory capabilities,the well-known traditional cryptographic techniques cannot be simply transferred to WSNs without adapting them.

3.1 Symmetric Cryptography

Symmetric encryption(also called as secret-key cryptography) uses a single secret key for both encryption and decryption.



3.1.1 Symmetric -Key Cryptography

This key has to be kept secret in the network, which can be quite hard in the exposed environment where WSNs are usedto achieve the security requirements, several researchers have focused on evaluating crypto graphical algorithms in WSNs and proposing energy efficient ciphers. Symmetric key algorithms are much faster computationally than asymmetric algorithms as the encryption process is less complicated. Examples are AES,3DES etc.

We first focus on Symmetric Cryptography due to the assumption that Symmetric cryptography has a higher effectiveness and require less energy , in contrast to public key cryptography.

According to [6] public key is used in some applications for secure communications eg.SSL(Secure Socket Layer) and IPSec standards both use it for their key agreement protocols.

But it consumes more energy and it is more expensive as compared to symmetric key has given a reason that public key consumes more energy due to great deal of computation andprocessing involved ,which makes it more energy consumptive as compared to symmetric key technique e.g. a single public key operation can consume same amount of time and energy as encrypting tens of megabits using a secret key cipher.

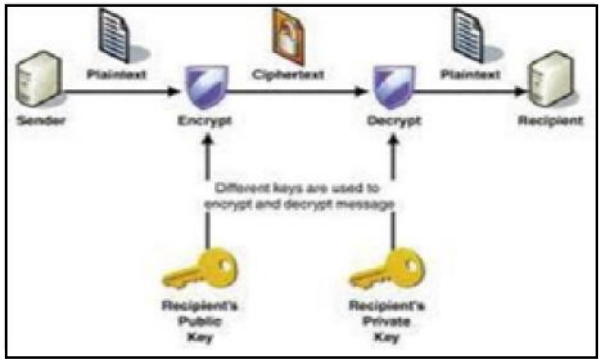
According to [8],the more consumption of computational resources of public key techniques is due to the fact that it uses two keys. One of which is public and is used for encryption ,and everyone can encrypt a message with it and other is private on which only decryption takes place and both the keys has a mathematical link, the private key can be derived from a public key.In order to protect it from attacker the derivation of private key from public is made difficult as possible like taking factor of a large number which makes it impossible computationally.Hence,it shows that more computation is involved in asymmetric key techniques thus we can say that symmetric key is better to choose for WSN.

Public Key cryptography is not only expensive in computation but also it is more expensive in communication as compared to symmetric key cryptography.According to [10] to send a public key from one node to another, at least 1024 bits required to be sent if the private key is 1024 bits long.

Two types of symmetric ciphers are used: block ciphers that work on blocks of a specific length and stream ciphers that work bitwise on data. A stream cipher can be seen as a block cipher with a block length of 1 bit.

3.2 Asymmetric Cryptography

Asymmetric encryption (also called public-key cryptography) uses two related keys (public and private) for data encryption and decryption, and takes away the security risk of key sharing. The private key is never exposed.



3.2. 1 Asymmetric Key Cryptography

A message that is encrypted by using the public key can only be decrypted by applying the same algorithm and using the matching private key. Likewise, a message that is encrypted by using the private key can only be decrypted by using the matching public key. Examples are RSA,ECC etc.

Public key Cryptography was omitted from the use in WSN because of its great consumption of energy and bandwidth which was very crucial in sensor network. Now a days a sensor become powerful in terms of CPU and memory power so, recently there has been a change in the research community from symmetric key cryptography to public key cryptography. Also symmetric key does not scale well as the number of nodes grows[15].

Arazi et al. [16] describe the efficiency of public-key cryptography for WSNs and the corresponding issues that need to be considered. Particularly, ECC is highlighted as suitable technique forWSN which provides a good trade-off between key size and security.

3.3 RSA Algorithm

A method to implement a public key cryptosystem whose security is based on the difficulty of factoring large prime numbers was proposed in [20].RSA stands for Ron Rivest, Adi Shamir and Leonard Adelman, who first publicly described the algorithm in 1977. Through this technique it is possible to encrypt data and create digital signatures. It was so successful that today RSA public key algorithm is the most widely used in the world.

3.3.1 Key Generation

Choose two distinct prime numbers, p and q. Compute modulus n = pq
Compute phi, ц= (p - 1)(q - 1) where ц is Euler's Totient Function.
Select public exponent e such that 1 < e < ц and gcd (e, ц) = 1
Compute private exponent d = e -1mod ц
Public key is {n, e}, private key is d
Encryption: c = me (mod n).
Decryption: m = cd(mod n).
Digital signature: s = H(m)d mod n, Verification: m' = se mod n, if m' = H(m) signature is correct. H is a publicly known hash function.

3.4 ECC (Elliptic Curve Cryptography)[21]

This algorithm is mainly depend on the algebraicstructure of elliptic curves.The difficulty in problemis ,the size of the elliptic curve. The primary benefit promised by ECC is a smaller key size, reducing storage and transmission requirements-i.e., that an elliptic curve group could provide the same level of security afforded by an RSA-based system with a large modulus and correspondingly larger key-e.g., a 256bit ECC public key should provide comparable security to a 3072bit RSA public key(see #Key sizes). For current cryptographic purposes, an elliptic curve is a plane curve which consists of the points satisfying the equation: y2=x3+ax+b, Compared to RSA,ECC has small key size, low memory usage etc. Hence it has attracted attention as a security solution for wireless networks [22].

3.5 Hybrid Cryptography

Symmetric key algorithm has a disadvantage of keydistribution[23] and asymmetric algorithm needmuch computation so the power of the sensor iswasted in it[23] and it is not feasible to use as poweris wasted then sensor will be of no useThus the algorithm which combines both thealgorithm i.e. asymmetric and symmetric so theadvantages of both the algorithm can be utilized in it. A hybrid cryptosystem is a protocol using multiple ciphers of different types together, each to it's best advantage. One common approach is to generate a random secret key for a symmetric cipher, and then encrypt this key via an asymmetric cipher using the recipient's public key. The message itself is then encrypted using the symmetric cipher and the secret key. Both the encrypted secret key and the encrypted message are then sent to the recipient.

The recipient decrypts the secret key first, using his/her own private key, and then uses that key to decrypt

the message. This is basically the approach used in PGP. Some of the hybrid algorithm like DHA+ECC[24] is described in detail.

4. PUBLIC-KEY CRYPTOSYSTEM

A public key cryptosystem employs a pair of different but associated keys. One of these keys is released to the public while the other, the private key, is known only to its owner. It is designed to be computationally intractable to calculate a private key from its associated public key; In other words, it is believed that any attempt to compute this key will fail during the lifetime of the network or the duration of an operation even when up-to-date technology and equipment are used.

With a public key cryptosystem, the sender can encrypt a message using the receiver's public key without needing to know the private key of the receiver. Therefore, they are suitable for communication among the general public.

Today, three types of systems, classified according to the mathematical problem on which they are based, are generally considered both secure and efficient. They are classified as follows:

x The integer factorization systems. x The discrete logarithm systems. x The elliptic curve discrete logarithm systems.

A. Public-Key Encryption (PKE)

A PKE is a triple of PPT algorithms

E = (G; E;

+ where:

G is the key-generation algorithm. G (1^k) outputs (PK; SK; Mk), where SK is the Secret key, PK is the public-key, and Mk is the message space associated with the PK/SK-pair. Here k is an integer usually called the security parameter, which determines the security level.

E is the encryption algorithm. For any m•Mk, E outputs cm••r E(m;PK) the encryption of m. c is called the cipher text. We sometimes also write

E(m;PK) as EPK(m), or E(m;r;PK)
and EPK (m;r), when we want to emphasize the randomness r used by E.

c) D is the decryption algorithm.

D(c;SK)••or m~•{invalid }* M is called the decrypted message. We also sometimes denote D(c;SK)
as DSK(c) and remark that usually D is deterministic.

We require the correctness property, i.e. everybody behaves as assumed:

m•Mk , m~ m, that is DSK(EPK(m)) m

Let us check that RSA satisfies the above definition. Notice, both E and D are deterministic.

G (1^k) corresponds to the following algorithm: p and q are random primes of k bits, n=pq; em••r =*M(n), d e^1 modM(n), Mk =n*.

Set PK= (n, e), SK=d.

- c E(m;(n,e)) me mod n.
- m~ D(c;(d,n)) cd mod n.

More generally, we could construct a PKE from any TDP. Suppose we have a TDP *f* with trap-door information tk and algorithm I for inversion. Here is the induced PKE:

- G(1^k)••or (*f* , tk , ^0,1'k), and *f* is the PK and trapdoor tk is the SK.
- E(m;PK) *f* (m).
- D(m;SK) I (c, tk).

4.1 RSA Cryptosystem

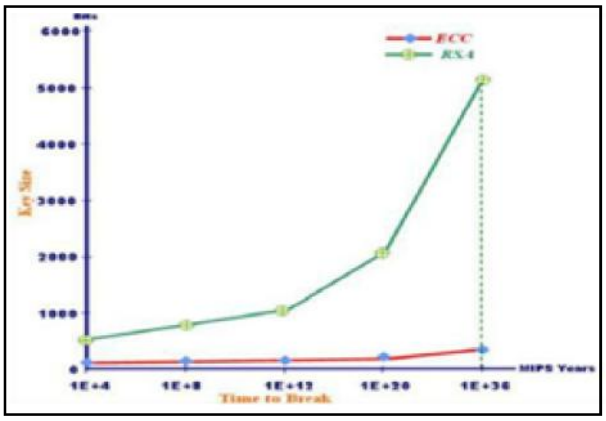
The RSA-system is based on the difficulty of factoring, P = C = Z/nZ for an integer n = p.q, where n (the modulus) is known to everybody, but the prime factors p, q are known only to receiver. We need in practice p and q to be very large. We take K to be the set of positive integers relatively prime to lcm(pi1, qi1). The encryption key e•K is known to everyone, but the decryption key d •K is known only to receiver. Then sender encrypts:

E : PuK ••oC, E(a,e) mod a^e n
To decrypt the cipher text, receiver:
D :C uK ••oP, D(b,e) mod b^d n
where ed { 1(mod lcm(p-1,q-1)) .

C. Elliptic Curve Cryptosystem (ECC) Elliptic curves are an algebraic structure, and their use for cryptography was first mentioned in [14] and [15]. They feature properties which allow the setup of a problem similar to the well known discrete logarithm problem of finite fields - also known as Galois fields (GF).

The subsequent section gives a brief and rough mathematical background to understand our implementation.

In recent years, ECC has attracted much attention as the security solutions for wireless networks due to the small key size and low computational overhead.



Comparison of Security Levels

Compares the time required to break ECC with the time required to break RSA for various modules sizes and using the best general algorithms known. The running times are computed in MIPS years.

From Fig. 1, we see that to achieve reasonable security, RSA should employ 1024-bit module, while a 160-bit module should be sufficient for ECC. Moreover, the security gap between the systems increases dramatically as the module sizes increases. For example, 160-bit ECC offers the comparable security to 1024-bit RSA.

ECC includes key agreement, encryption, and digital signature algorithms. The key distribution algorithm is used to share a secret key, the encryption algorithm enables confidential communication, and the digital signature algorithm is used to authenticate the signer and validate the integrity of the message.

5. CONCLUSION

The wireless sensor networks continue to grow and become widely used in many applications. So, the need for security becomes vital. However, the wireless sensor network suffers from many constraints such as limited energy, processing capability, and storage capacity, etc. There are many ways to provide security, one is cryptography. Selecting the appropriate cryptography method for sensor nodes is fundamental to provide security services in WSNs. Public Key based cryptographic schemes were introduced to remove the drawbacks of symmetric based approaches.We have compared two schemes in this paper ECC, and RSA and found out that ECC is more advantageous compared to RSA,due to low memory usage,low CPU consumption and shorter key size compared to RSA.ECC 160 bits is two times better than RSA 1024 bits when code size and power consumption are the factors of consideration. Tests were performed in 8051 and AVR platforms as in[25].ECC 160 bits use four times less energy than RSA 1024 bits in Mica2dot as in[26].Recently a new scheme called Multivariate Quadratic Almost Group was proposed which showed significant improvements over RSA and ECC.

REFERENCES

[1] D. Djenouri and L. Khelladi, "A Survey of Security Issues in Mobile Ad Hoc and Sensor Networks", IEEE Communication Surveys and Tutorials, Vol.7, No.4, December 2005, pp.2-28.

[2] G. Gaubatz, J-P. Kaps and B. Sunar, "Public Key Cryptography in Sensor Networks Revisited", 1st European Workshop on Security in Ad-Hoc and Sensor Networks (ESAS 2004), Lecture Notes in Computer Science, Springer, Heidelberg, Vol.3313, August, 2004, pp. 2-18.

[3] A. Perrig, R. Szewczyk, V. Wen, D. Culler and J. D. Tygar, "SPINS: Security Protocols for Sensor Networks", Wireless Networks, Vol.8, No.5, September 2002, pp. 521–534.

[4] A. Baggio, "Wireless Sensor Networks in Precision Agriculture", in ACM Workshop on Real-World Wireless Sensor Networks (REALWSN 2005), Stockholm, Sweden, June 2005.

[5] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler and J. Anderson, "Wireless Sensor Networks for Habitat Monitoring", in First ACM Workshop on Wireless Sensor Networks and Applications, Atlanta, GA, USA, September 2002.

[6] G. Fuchs, S.Truchat and F. Dressler, "Distributed Software Management in Sensor Networks using Profiling Techniques", in1stIEEE/ACM International Conference on Communication System Software and Middleware (IEEE COMSWARE 2006): 1st International Workshop on Software for Sensor Networks (SensorWare 2006), New Dehli, India, January 2006.

[7] W. Zhang, G. Cao, "Group Rekeying for Filtering False Data in Sensor Networks: A Predistribution and Local Collaboration-Based Approach", in 24th IEEE Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE INFOCOM 2005), March 2005, pp.503-514.

[8] K. Piotrowski, P. Langendoerfer, S. Peter, "How Public Key Cryptography Influences Wireless Sensor Node Lifetime", Proceedings of the fourth ACM workshop on Security of ad hoc and sensor networks, USA, 2006, pp.169-176.

[9] Y. W. Law, "Key Management and Link-Layer Security of WSN", Ph.D. Thesis,University of Twente, Netherland, 2005.

[10] A.Perrig, R.Szewczyk, V.Wen, D. Culler and J. D.Tygar, "SPINS: Security Protocols for Sensor Networks", in Proceedings of the 7th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom), Rome, Italy, July 2001, pp.189-199.

[11] S.Zhu, S. Setia and S. Jajodia, "LEAP:Efficient Security Mechanisms for Large-Scale Distributed Sensor Networks", In the Proceedings of the 10th ACM conference on Computer and Communications Security, 2003.

[12] J.P.Walters,Zh.Liang, W.Shi and V.Chaudhary, "Security in Distributed, Grid, and Pervasive Computing", Chapter 17,CRC Press, 2006.

[13] R. B. Ghazali, "Security in WSN in Enhance AODV Routing", Masters Thesis, Faculty of Electrical Engineering, University Technology Malaysia, 2006.

[14] N. Koblitz, "Elliptic Curve Cryptosystems", Mathematics of Computation, Vol. 48, 1987.

[15] V.S. Miller, "Use of Elliptic Curves in Cryptography", Advances in Cryptology CRYPTO 85, 1986.

[16] A. J. Menezes, P. C. Van Oorschot, S. A. Vanstone, "Handbook of Applied Cryptography", CRC Press, 1996.

[17] W. Du, J. Deng, Y. S. Han, Shigang Chen and P.K.Varshney, "A Key Management Scheme for Wireless Sensor Networks Using Deployment Knowledge", IEEE INFOCOM 2004.

[18] B. C. Neuman and T. Tso, "Kerberos: An Authentication Service for Computer Networks", IEEE Communications, Vol. 32, No.9, September 1994, pp. 33-38.

A Basic Paper on Data Security and HADOOP File System

R.Deepa and S.Vaishnavi
Department of Information Technology, PSG College of Arts and Science, Coimbatore - 641 014 , Tamil Nadu

Abstract

Hadoop is a Java Software framework which supports data and time period of several applications and Developed in open source license. It enables applications to work with thousands of nodes and Petabytes of data. The two major parts of hadoop are HDFS: Hadoop's own File System. The scale is designed to for storage of petabytes and runs on top of the file systems of the underlying operating systems.

1. INTRODUCTION

Hadoop was developed from GFS (Google File System) [2, 3] and Map Reduce papers published by Google in 2003 and 2004 respectively. Hadoop is a framework of tools and it is implemented in Java. It supports applications which is running on big data.

1.1 Project Plan

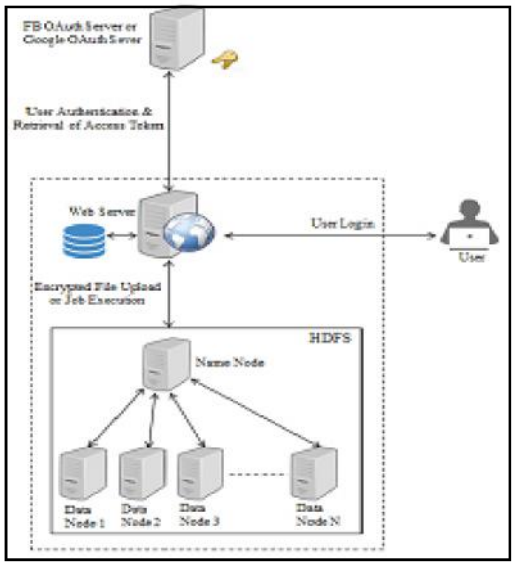
Hadoop is designed without considering security of data. HDFS is in plaintext of Data Storage. This data is p accessed by users who are unauthorized. The method for securing method data is needed. So we are developing this highly secured system for Hadoop Distributed File System.

1.2 Requirement of Project

Hadoop is generally executing in big clusters or might be in an open cloud administration. Amazon, Yahoo, Google, and so on are such open cloud where numerous clients can run their jobs utilizing Elastic MapReduce and distributed storage provided by Hadoop. It is key to execute the security of client information in such systems. Web produces expansive measure of information consistently. It incorporate the organized information rate on web is around 32% and unstructured information is 63%. Additionally the volume of advanced substance on web grows up to more than 2.7ZB in 2012 which is 48% more from 2011 and now soaring towards more than 8ZB by 2015. Each industry and business associations are has a critical information about various item, generation and its business sector review which is a major information advantageous for efficiency development.SThe files in Hadoop distributed file system (HDFS) are divided into multiple blocks and replicated to other DataNodes(by default 2 nodes) to ensure high data availability and durability in case of failure of execution of job(parallel application in Hadoop environment). Originally Hadoop clusters have two types of node operating as master-salve or master-worker pattern . NameNode is a master node and DataNodes are workers nodes in HDFS. Data nodes are the nodes where actual file(part of file on a node) is stored. However NameNode contains information about where the different file blocks are located but it is not persistent, when system starts block may changes one DataNode to another DataNode but it report to NameNode or client who submit the MapReduce job or owner of Data periodically . Client gets list of data nodes where file blocks reside & then communicate with Data nodes only. NameNode contains only metadata.

2. DIAGRAM OF HADOOP

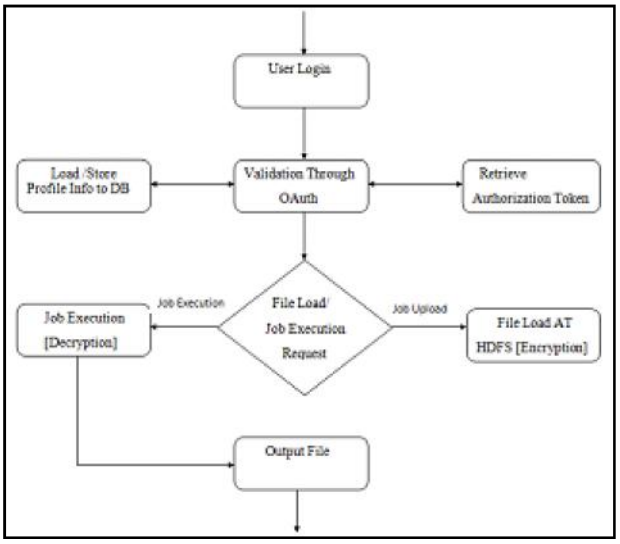
2.1 System Architecture



3. PROPOSED SYSTEM

We have proposed new technique for securing data at HDFS by analyzing all techniques previously mentioned. It is actualized by utilizing Real Time Encryption Algorithm and OAuth (called Open Standard for Authorization). OAuth 2.0 is an Open Authentication Protocol that is used for authentication and authorization of client in conventional client-server model. In the traditional client-server model, the customer solicitations to an entrance secured asset on the server by verifying itself utilizing the asset proprietor's international ID. In order to give third-party applications access to restricted resources .In proposed system, to authenticate user we have used OAuth 2.0, which returns unique token for each user who attempts successful login. The token returned by OAuth server utilized as a part of encryption strategy so it gives information privacy and integrity to the user data. The files are encrypted before load to HDFS and decrypted when job execution is in progress [1]. The Real Time Encryption Algorithm utilizes the OAuth token as key and Encrypt data (uploaded by user).

3.1 Flow Diagram



4. CONCLUSION

Hadoop has been very much solutions which is effective for dealing with companies data which has petabytes. It solves many problems in industries which relates in huge data of Distributed and Management System. It is open source which is widely used by several companies.

Nowadays, Buzzword is Big data from news article to media, YouTube from blog discussions to science projects. The area is, not surprisingly, computer science the topic such as engineering, mathematics, business and also social and decision sciences. There is also much to do in developing a professional approach to big data mining.

REFERENCES

[1] Seonyoung Park and Youngseok Lee, "Secure Hadoopwith Encrypted HDFS", Springer-Verlag Berlin Heidelberg in 2013.
[2] J. Dean and S.Ghemawat, "MapReduce: Simplified DataProcessing on Large Cluster", In:OSDI (2004).
[3] T. White, "Hadoop: The Definitive Guide", 1st edn.OReilly Media (2009)
[4] Jason Cohen and Dr. Subatra Acharya Towards.
[5] "Big Data Security: The Evolution of Hadoops Security Model", Posted by Kevin T. Smith on Aug 14, 2013
[6] Hadoop, http://hadoop.apache.org/

Emotion Recognition Using Affective Sound Stimulation through Heart Rate Variability

S. Suganya and J C Miraclin Joyce Pamila
Department of Computer Science and Engineering,
Government College of Technology, Coimbatore- 641 013, Tamil Nadu
E-mail:ssugumoni@gmail.com

Abstract

This paper reports on the emotional states of different arousal levels, elicited by affective sounds can be effectively recognized by means of estimates of Autonomic Nervous System (ANS) dynamics. Specifically, emotional states are modeled as a combination of arousal and valence dimensions according to well-known Russell s circumplex model of affect, whereas the ANS dynamics is estimated through standard and nonlinear features of Heart rate variability (HRV) exclusively, which is derived from the electrocardiogram (ECG) . Standard methods as well as nonlinear dynamic techniques were used to extract sets of features from the collected peripheral physiological signals using Kubios HRV software. In addition, Lagged Poincare Plots of the HRV series and then non-linear approach based on Hurst feature were also taken into account for nonlinear feature extraction. The non-linear feature „Hurst component was computed using Rescaled Range Statistics (RRS) and Finite Variance Scaling (FVS). The affective sounds were gathered from the International Affective Digitized Sound System (IADS) and grouped into four different levels of arousal (intensity of the sound) and two levels of valence (unpleasant and pleasant sounds). A group of 27 healthy volunteers were administered with these standardized stimuli while ECG signals were continuously monitored. Using Quadratic Discriminant Classifier (QDC) algorithm, tested through Leave-One-Subject-Out (LOSO) procedure, was able to achieve recognition accuracy for valence dimension and arousal dimension.

Keywords: *Emotion Recognition, Russell s Circumplex model of affect, Autonomic Nervous System, International Affective Digitized Sound System (IADS), Heart rate variability, Nonlinear Feature Extraction, Poincare plot, Quadratic Discriminant Classifier (QDC).*

1. INTRODUCTION

Affective Computing is a relatively new and fast growing research area which combines knowledge in the field of computer science, psycho-physiology, biomedical engineering and artificial intelligence. In general, an emotion recognition system is designed to be effective for a specific kind of affective stimulus and it is built on a specific model of emotion which has to be characterized by processing one or more physiological and behavioral signs. In this study, the Russell s Circumplex model of affect [1], [2], which is one of the most widely used model of emotions, elicited by affective sounds. More specifically, considering such a model, each emotion is viewed as a combination of two affective dimensions: arousal (ranging from low to high strength of the emotion) and valence (ranging from unpleasant to pleasant).

Concerning the physiological signals to be taken into input for emotion recognition, ECG-derived signals, which mainly refer to the analysis of Heart rate variability (HRV), have been extensively proposed in this literature. For instance, using IADS (International Affective Digitized Sound System) stimuli, significant changes in facial ElectroMyoGram (EMG), Electroencephalography (EEG) activities were reviewed [3] and along with Autonomic Nervous System (ANS) derived signals [4],[5],[6],[7].

Majority of the affective computing studies, based on ANS dynamics, in this study we have done an effective emotion recognition system based on the measures derived from the HRV. Also regarding the affective emotion elicitation, a wide range of affective elicitation methods have been done in this literature computer game interfaces [8], images [6] as spoken words [9], music [10], real experiences [11], film clips [12]. Specifically, we selected the standardized acoustic sound stimulation gathered from the IADS and it is already scored in terms of arousal levels and valence levels, allowing us to determine the different arousal levels and valence levels of the elicitation. Moreover, use of standardized stimuli allows replicating studies in a more advanced fashion and making easier the comparison of the results with the future related works.

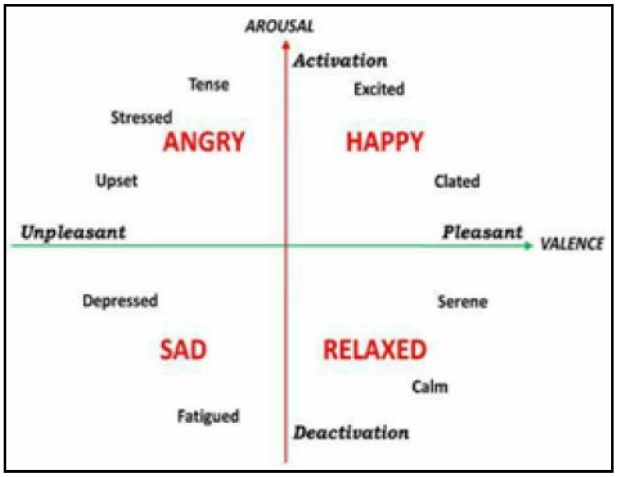


Figure 1: A graphical representation of the Russell s circumplex model of affect in terms of arousal and valence dimensions: the horizontal axis representing the valence or pleasant dimensions and the vertical axis represent the arousal or activation dimensions.

In this work, we first provide a detailed characterization of the HRV dynamics, and then we select affective sound stimuli from IADS having different levels of arousal and valence, including the neutral level. Then features are derived from standard measures, which are defined in terms of time and frequency domain, from the HRV [14] and then non-linear measures, which are defined in terms of phase space domain analysis.

As a novelty of this work, Lagged Poincare Plots (LPP) were taken into account, which is a non-linear technique widely used for the study of HRV dynamics. Finally those HRV measures showing statistical significant differences between arousal levels and valence levels are considered as an input of Quadratic Discriminant Classifier (QDC) used for the automatic emotion recognition system.

2. MATERIALS AND METHODS

2.1 Subjects Recruitment, Experimental Protocol and Acquisition Set-up

Twenty-seven healthy subjects, aged from 25 to 35, voluntarily participated in the experiment. According to self-report questionnaires, none of them had a history of injury of the auditory canal or partial or full incapability of hearing and then none of them suffered from any cardiovascular, mental or chronic disease. Volunteers were informed about the experimental protocol and about the purpose of the study, but they were not informed about the arousal and valence levels that they would have been listened to. During the experimental protocol, volunteers were seated in a comfortable chair listening to the IADS stimuli through headphones, keeping their eyes closed in order to avoid any kind of visual interference.

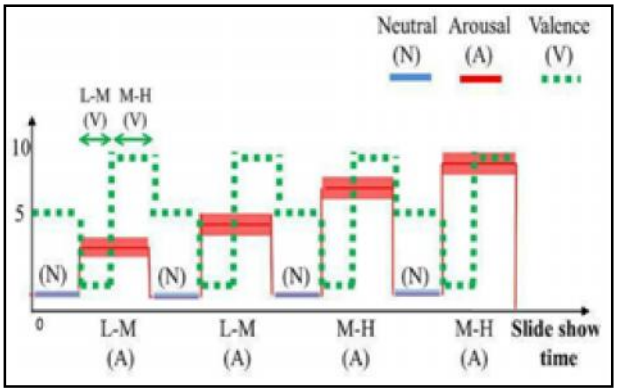


Figure 2: Timeline sequence scheme of the experimental protocol in terms of arousal and valence levels. The y-axis relates to the official IADS score, whereas the x-axis relates to the time. The neutral sessions, which are marked with blue lines, alternate with arousal ones, which are marked with red staircases. Along the time, the red line follows the four arousal sessions having increasing intensity of activation dimension. The dotted green line indicates the valence levels, used to distinguish the low-medium (L-M) and the medium-high (M-H) level with an arousing session. The yellow line relates to the resting state. The neutral sessions are characterized by lowest arousal and medium valence scores.

The affective elicitation protocol was divided into 10 sessions: after an initial resting session of 5 minutes, four arousal sessions alternated with neutral session (Figure 2). The four arousal levels had increasing scores (Table I). Within each arousing levels the standardized acoustic stimuli were selected to have Low-Medium (L-M) for negative valence levels and Medium-High (M-H) for positive valence levels. The neutral session had duration of 1 minute and 28 seconds. The four arousal sessions had duration of 3 minutes and 30 seconds, 3 minutes and 40 seconds, 4 minutes, 5 minutes and 20 seconds respectively. This experimental protocol was approved by the local ethical committee of the University of Pisa, Italy.

Throughout the whole experimental session, the ECG signal was continuously monitored, following the Einthoven triangle configuration, by means of a hardware

module i.e., the ECG100C Electrocardiogram Amplifier from BIOPAC Inc. with a sampling rate of 500 Hz. From the ECG signal, HRV series is extracted, which refers to the variation of the time intervals between consecutive heartbeats identified with the R-waves (RR intervals). Therefore, to obtain HRV series, QRS complex detection algorithm was used, i.e., the automatic algorithm developed by Pan-Tompkins [15]

Table 1 Rating of IADS Sounds Used in this Work

Session	Number of Sounds	Valence Rating	Valence Range	Arousal Rating	Arousal Range
Neutral	8	5.915±0.68	4.34÷6.44	3.47±0.175	2.88÷3.93
Arousal 1	19	/	3.54÷7.51	4.60±0.21	4.03÷4.97
Arousal 2	19	/	2.46÷7.78	5.42±0.22	5.00÷5.89
Arousal 3	26	/	2.04÷7.90	6.48±0.25	6.00÷6.99
Arousal 4	20	/	1.57÷7.67	7.32±0.22	7.03÷8.16

Ratings are expressed as median and its absolute deviation.

2.2 Methodology of Signal Processing

A detailed description of the methodology of signal processing for the extraction of standard and non-linear features follows below.

2.2.1 Standard HRV Measures

Standard HRV Measures refers to the extraction of parameters defined in terms of time and frequency domain [4], [14].

Concerning the time domain measures, we calculated the following HRV features:

- the mean value (RR mean)
- the standard deviation (RR std)
- the standard deviation of NN intervals (Normal
- to-Normal intervals) (SDNN)
- the square root of the mean of the sum of the square of differences between Subsequent NN intervals
- the number of successive differences of intervals which differ by more than 50ms, expressed as a percentage of the total number of heartbeats analyzed
- the integral of the probability density distribution (i.e., number of all NN intervals) divided by the maximum of the probability density distribution (HRV triangular index)
- the triangular interpolation of NN interval histogram i.e., baseline width of the distribution measured as a base of a triangle approximating the NN interval distribution (TINN)

Concerning the frequency domain analysis, we calculated the following features from Power Spectral Density (PSD) analysis by using the Welch s periodogram method. Three spectral bands of the Power Spectral Density were identified: Very Low Frequency (VLF) with spectral components below 0.04 Hz, Low Frequency (LF) ranging frequencies between 0.04 Hz and 0.15 Hz, High Frequency (HF) ranging frequencies between 0.15 Hz to 0.4 Hz. Following features were computed for each of the three frequency bands:

- the power calculated within the VLF, LF and HF bands
- the frequencies having maximum magnitude (VLF peak, LF peak and HF peak)?
- the power expressed as percentage of the total power (VLF power%, LF power % and HF power%)
- the power normalized to the sum of the LF and HF power
- the LF/HF power ratio

2.2.2 Nonlinear HRV Measures

In this study, non-linear features are extracted using the following methods: the Approximate Entropy (ApEn), the Detrended Fluctuation Analysis (DFA) Lagged Poincare Plots (LPP).

i. Approximate Entropy

Approximate Entropy is a measure of the unpredictability in the time series. A lower value of ApEn corresponds to a repetitive trend, whereas higher is the ApEn value as complex is the signal.

ii. Detrended Fluctuation Analysis (DFA)

Detrended Fluctuation Analysis is a method for determining the statistical self-affinity of a signal. DFA is related to the measures based upon spectral techniques such as auto correlation and Fourier transform and is useful in revealing the extent of long-range correlation in time series [16]. In DFA the correlations are divided into long-term fluctuations and short-term fluctuations, where the short-term fluctuations of the signal are characterized by the parameter $\alpha 1$ and the long-term fluctuations are characterized by the parameter $\alpha 2$.

iii. Lagged Poincare Plots (LPP)

This technique is used to quantify the fluctuation of dynamics of the time series through a graphic representation (scatter plot of RR intervals) where each current RRn interval is mapped as a function of previous interval. In this work we also used Lagged Poincare Plots, a scatter plot of RR_n and RR_{n+M}, with $1< M < 10$

Figure 3: Overall block scheme of the proposed emotion recognition system. The ECG is preprocessed in order to extract the RR interval series. According to the experimental protocol, standard and nonlinear features are extracted and, then, selected through statistical analysis. After a normalisation step, Quadratic Discriminant Classifier algorithms are engaged to perform pattern recognition by adopting a Leave-One-Subject-Out procedure.

The quantitative analysis from the graph is calculated using the following measures:

- SD1: the standard deviation related to the points that are Perpendicular to the line-of-identity $RR_{n+M} = RR_n$ [17]. It describes briefly about the HRV short-term variability.
- SD2: the standard deviation that describes the long-term dynamics and measures the dispersion of the points along the identity line.
- SD12 (SD12 = SD1/SD2), the ratio between SD1 and SD2. It measures the balance between the HRV long-term variability and short-term variability [17].
- S (S = PSD1SD2), the area of an imaginary ellipse with axes SD1 and SD2 [16][18].
- An approximate relation indicating the variance of the whole HRV series [19].

The non-linear feature „Hurst component was computed using two methods Rescaled Range Statistics (RRS) and Finite Variance Scaling (FVS). Then new Hurst features were computed by combining Rescaled Range Statistics (RRS) and Finite Variance Scaling (FVS) with Higher Order Statistics (HOS). These features convey the information related to the properties such as similarity, predictability, reliability and then sensitivity of the signal.

2.2.3 Statistical Analysis and Pattern Recognition

In this study, we use statistical analysis and pattern recognition methodologies to automatically recognize the emotional responses of the subject. This study has been implemented using the following Leave-One-Subject-out-Procedure (LOSO): we applied the feature selection procedure through statistical analysis and the normalization of the features on a training set made by N-1 subject (where N is the total number of volunteers) to automatically recognize the emotional responses of the subject N_{th}. This LOSO procedure was iterated N times. Three statistical tests were performed: Kolmogorov-Smirnov test, Friedman test, Wilcoxon signed rank- test. Kolmogorov-Smirnov test were carried out in order to check whether data were normally distributed. In case of non-normal distribution, as considering the paired data, a Friedman test [20], was used to test the null hypothesis that no difference exists among arousal and valence sessions. Wilcoxon signed-rank test was applied to find out the significant differences between each arousal session, valence session and neutral one. Standard and nonlinear measures showing a probability of null-hypothesis as ($p<0.05$) were used as input of a Quadratic Discriminant Classifier (QDC) which was validated through LOSO procedure [18],[19]. The classification results were expressed as recognition accuracy in terms of confusion matrices.

3. EXPERIMENTAL RESULTS

Concerning the HRV standard measures, throughout the iterations tested through Leave-One-Subject-Out (LOSO) rocedure, significant differences were found on the RRmean, RMSSD, RRstd and TINN. Then, concerning the HRV features defined in terms of frequency domain, significant differences were found on LF power%, LF power nu, HF power, HF power%, HF power nu and LF/HF power ratio. Then, concerning the HRV non-linear features, significant differences were found on DFA$\alpha 1$, DFA$\alpha 2$, ApEn and SD1, SD2, SDRR

and S (Extracted using Poincare Plots). Of note, LPP derived measures showed a significant p-value($p<0.05$) among the arousal and neutral sessions. In particular heartbeat dynamics, shown as a function of the arousal level, is related to the SD12 measures whose Lagged Poincare Plots are reported in Figure 4. It is clear to notice that the degree of separation between the arousing and the neutral sessions increases according to the degree of the arousal level of the acoustic stimuli. Of note, neutral sessions were not taken into account for the automatic classification system because each auditory stimuli served for the normalization before the classification. Experimental results demonstrated that using a Quadratic Discriminant Classifier (QDC), validated through Leave-One-Subject-Out procedure, was able to achieve recognition accuracy for both arousal and valence dimensions. As expected, highest accuracy was obtained while discerning the arousal and valence levels.

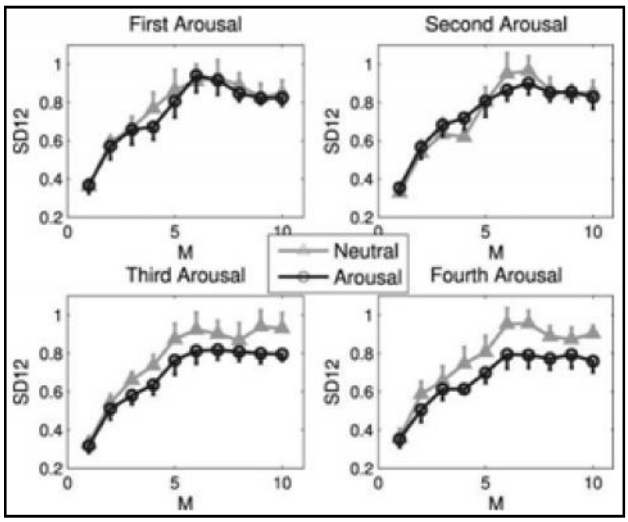


Figure 4: SD12 values as a function of the M lags among the four arousal sessions and corresponding previous neutral sessions. The values are expressed in terms of medians and MAD.

4. CONCLUSION AND DISCUSSION

In conclusion, we presented a novel approach to automatically recognize the emotional states based on the signals of HRV standard and nonlinear features. Emotions are expressed in terms of arousal levels and valence levels according to the Russell s circumplex model of affect. This study also reveals that Autonomic Nervous System measures such as HRV mean value, standard deviation, RMSDD, triangular index, spectral band measures, Approximate entropy, Detrended fluctuation analysis, standard deviation of the Poincare plot to automatically recognize the emotional states of a person induced by the affective sounds. In this study [7], many classification algorithms were used such as Linear Discriminant Classifier (LDC), Multilayer Perceptron (MLP), k-Nearest Neighbor, among these Quadratic Discriminant Classifier showed the highest recognition accuracy in terms of both arousal and valence dimensions. More in general, number of previous research works highlighted the crucial role of non-linear dynamics for automatic classification of emotion recognition [10],[16]. Future work will focus on the investigation of improving the accuracy of arousal and valence dimensions by adding the features of the physiological signals. band measures, Approximate entropy, Detrended fluctuation analysis, standard deviation of the Poincare plot to automatically recognize the emotional states of a person induced by the affective sounds. In this study [7], many classification algorithms were used such as Linear Discriminant Classifier (LDC), Multilayer Perceptron (MLP), k-Nearest Neighbor, among these Quadratic Discriminant Classifier showed the highest recognition accuracy in terms of both arousal and valence dimensions. More in general, number of previous research works highlighted the crucial role of non-linear dynamics for automatic classification of emotion recognition [10],[16]. Future work will focus on the investigation of improving the accuracy of arousal and valence dimensions by adding the features of the physiological signals.

REFERENCES

[1] J. A. Russell, “A Circumplex Model of Affect”, J.Personality Soc. Psychol., Vol. 39, No. 6, 1980, pp.1161.

[2] J. Posner, J. A. Russell, B. S.Peterson,”The Circumplex Model of Affect: An Integrative Approach to Affective Neuro Science, Cognitive Development and Psychopathology”, Develop.Psychopathol. Vol.17, No.03, 2005, pp. 715-734.

[3] M. M. Bradley and P. J. Lang, “Affective Reaction to Acoustic Stimuli”, Psychophysiology, Vol.37, No.02, 2000, pp.204-215.

[4] G. Valenza, A. Lanata, and E. P. Scilingo, “The Role of Nonlinear Dynamics Inaffective Valence and Arousal Recognition”, IEEE Trans. Affective Comput., Vol. 3, No. 2, Apr.-Jun. 2012, pp. 237-249.

[5] G. Valenza, M. Nardelli, A. Lanata, C. Gentili, G. Bertschy, R. Paradiso and E.P. Scilingo, "Wearable Monitoring For Mood Recognition In Bipolar Disorder Based On History-Dependent Long-Term Heart Rate Variability Analysis", IEEE J. Biomed. Health Informatics, Vol. 18, No. 5, Sep. 2014, pp. 1625-1635.

[6] K. H. Kim, S. Bang, and S. Kim, "Emotion Recognition System Using Short-Term Monitoring of Physiological Signals", Medical Biology Engineering Computing, Vol. 42, No. 3, 2004, pp. 419-427.

[7] J. Wagner, J. Kim, and E. Andre, "From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification", in Proc. IEEE Int. Conf. Multimedia Expo, 2005, pp. 940-943.

[8] J. Scheirer, R. Fernandez, J. Klein, and Picard,"Frustrating the User on Purpose: A Step Toward Building An Affective Computer," Interacting with Computing Vol. 14, No. 2, 2002, pp. 93-118.

[9] M. Ilves and V. Surakka, "Heart Rate Responses to Synthesized Affective Spoken Words", Adv. Human-Computing Interaction, 2012, pp. 14.

[10] Y. -H. Yang and H. H. Chen, Music Emotion Recognition. Boca Raton, FL, USA: CRC Press, 2011.

[11] L. Li and J.-h. Chen, "Emotion Recognition Using Physiological Signals," in Proc. Adv. Artificial Reality Tele-Existence, 2006, pp. 437-446.

[12] J. Rottenberg, R.D. Ray and J. J. Gross, "Emotion Elicitation Using Films," The Handbook of Emotion Elicitation and Assessment, 2007, pp.9-28.

[13] M. M. Bradley and P. J. Lang, "The International Affective Digitized Sounds : Affective Ratings of Sounds and Instruction Manual," University of Florida, Gainesville, FL, Tech. Rep. B-3, 2007.

Enhancement on the Performance Impact of Elliptic Curve Cryptography on DNSSEC Validation

R.Sangavi
Department of Information Technology, Kondu Engineering College, Perundurai, Erode- 638 060, Tamil Nadu
E-mail: sangaviravi95@gmail.com

Abstract

The name system (DNS) could be a core net infrastructure that interprets names to machine-readable data, such as scientific discipline addresses. Security flaws in DNS to a significant overhaul, with the introduction of the DNS security (DNSSEC)extensions. DNSSEC adds integrity and authenticity to the DNS by using digital signatures. DNSSEC, however, has its own issues. It availableness issues owing to packet fragmentation and could be a potent supply of distributed denial-of-service attacks. In earlier work, we have a tendency to argued that a lot of problems with DNSSEC stem from the selection of RSA as default signature algorithmic rule. A switch to alternatives supported elliptic curve cryptography (ECC) will resolve these problems. nevertheless shift to error correction code introduces a replacement problem: error correction code signature validation is way slower than RSA validation. Thus, shift DNSSEC to error correction code imposes a major additional burden on DNS resolvers, pushing load toward the edges of the network. Therefore, during this paper, we study the question: can shift DNSSEC to error correction code result in issues for DNS resolvers, or will they handle the additional load? To answer this question, we have a tendency to developed a model that accurately predicts however many signature validations DNS resolvers have to be compelled to perform. This allows USA to calculate the extra electronic equipment load error correction code imposes on a resolver. victimisation real-world measurements from four DNS resolvers and with two ASCII text file DNS implementations, we evaluate future eventualities wherever DNSSEC is universally deployed. Our results once and for all show that shift DNSSEC to error correction code signature schemes doesn't impose associate degree insurmountable load on DNS resolvers, even in worst case eventualities.

Keywords: *DNS, DNSSEC, Elliptic curve cryptography, ECDSA, EdDSA, ECC*

1. INTRODUCTION

The name System (DNS) is arguably one amongst the most crucial protocols on the net. Its main task is to translate human-readable names (such as www.utwente.nl')to computer code info (such as scientific discipline addresses). Over the past decade, the DNS has been undergoing a significant overhaul with the introduction of the DNS Security Extensions (DNSSEC). DNSSEC addresses a vital flaw within the DNS protocol: an absence of authenticity and integrity. This is often done the lack of trust within the original DNS protocol, it's not while not its own flaws.

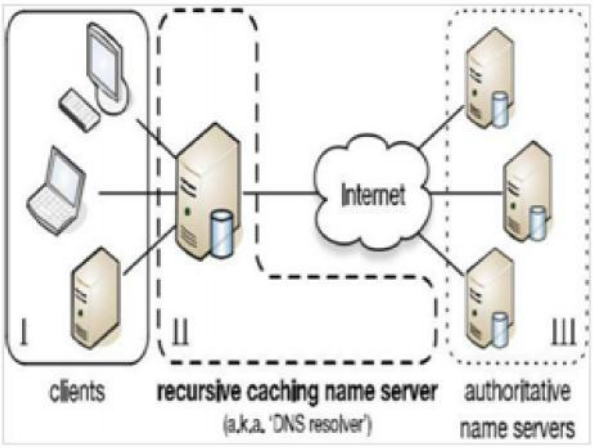
In earlier work, We have got shown that: •DNSSEC suffers from information processing fragmentation. As DNSSEC responses area unit larger than classic' DNS, due to the inclusion of digital signatures, they'll be fragmented at the information processing level. Up to 100% of DNS resolver on the net may not be ready to traumatize fragmented responses. This can have consequences for users of those resolvers. Resolving well-liked DNSSEC-signed domains, such as paypal.com, might incur a performance penalty and in the worst case the domain might even become unapproachable. •DNSSEC will be abused for potent distributed denial-of service attacks. As a result of DNS is prone to information processing address spoofing, and it will be abused in alleged amplification attacks. For classic' DNS the typical amplification issue is around 6Ч, however DNSSEC makes things a lot of worse, increasing the typical amplification to around50Ч[2].This means that by causation a hundred Mbit/s, attackers will mount associate degree attack of 5Gbits. The root reason for these problems is that the alternative of RSA as default signature algorithmic program for DNSSEC. We showed that various signature schemes supported elliptic curve cryptography (ECC) effectively address the main problems in DNSSEC represented higher than [3]. This can be as a result of code signatures are considerably smaller in size, resulting in smaller DNS responses. While change DNSSEC to ECC-based signature algorithms is

highly helpful and solves serious problems in DNSSEC, it introduces a brand new problem: validation of code signatures is associate degree order of magnitude slower than validation of the RSA signatures presently in wide spread use in DNSSEC. This may have consequences for the world DNS infrastructure.

Currently, using RSA, the foremost mainframe intensive operation in DNSSEC is that the language method. this can be performed at regular intervals by the DNS operators of signed domains. Validation of signatures is performed by algorithmic caching name servers(DNS resolvers'). Thus, a switch from RSA to ECC-based signatures imposes a major extra burden on DNS resolvers, effectively pushing the value of science operations in DNSSEC to the sides of the network.

The design of the DNS are often divided into three components as shown in Figure one. the primary half consists of client, shown on the left (I).Clients are usually have what's known as a stub resolver as a part of the package or an application such as an internet browser. The stub resolver performs DNS lookups on behalf of applications on the client. Stub resolvers area unit easy pieces of software system that source DNS lookups to a algorithmic caching name server, shown within the middle(II).The DNS name area may be a tree structure, beginning with the basis zone, followed by superior domains(such as.com,.net,..) one level down from the basis, and second-level domains (such as example.com) below that, and so on. Recursive caching name servers perform the particular DNS search through a method known as formula. throughout formula, they traverse the name area is from prime to bottom, act with authoritative name servers, shown on the correct(III).

Algorithmic caching subsequent clients causation constant question can receive the cached response till the TTL expires. Caching ensures that the expensive method of formula (in terms of network spherical trips) doesn't ought to be performed for each question. This paper studies the impact of DNSSEC on algorithmic caching name servers. These servers are a unit usually named as DNS resolvers'. A DNS resolver that validates the digital signatures employed in DNSSEC is then named as a validating DNS resolver .



2. LITERATURE SURVEY

2.1Check Repeat: A NEW METHOD OF MEASURING DNSSEC VALIDATING RESOLVERS

Y., Wessels, D., Larson, M., & Zhang, L. proposed a more and more authority DNS servers turn on DNS security extensions (DNSSEC), it becomes increasingly important to understand whether, and how many, DNS resolvers perform DNSSEC validation. In this paper presented a query-based measurement method, called Check-Repeat, to gauge the presence of DNSSEC validating resolvers. Utilizing the fact that most validating resolver implementations retry DNS queries with a different authority server if they receive a bad DNS response, Check-Repeat can identify validating resolvers by removing the signatures from regular DNS responses and observing whether a resolver retries DNS queries. Then analyze the Check-Repeat in different scenarios and our results showed that Check-Repeat can identify validating resolvers with a low error rate. Also cross-checked our measurement results with DNS query logs from .COM and .NET domains, and confirmed that the resolvers measured in this research can account for more than 60% of DNS queries in the Internet.

2.2 DNSSEC MISCONFIGURATIONS: HOW INCORRECTLY CONFIGURED SECURITY LEADS TO UNREACHABILITY

Van Adrichem, N.L., Lъa, A.R., Wang, X., Wasif, M., Fatturrahman, F., & Kuipers, F.A.,offered protection against spoofing of DNS data by providing authentication of its origin, ensuring integrity and giving a way to authenticate denial of existence by using public-key cryptography. Where the relevance of securing a technology as crucial to the Internet as DNS is obvious,

the DNSSEC implementation increases the complexity of the deployed DNS infrastructure, which may manifest in misconfiguration. A misconfiguration not only leads to silently losing the expected security, but might result in Internet users being unable to access the network, creating an undesired unreachability problem. In this paper, measure and analyze the misconfigurations for domains in four zones (.bg, .br, .co and .se). Furthermore, classify these misconfigurations into several categories and provide an explanation for their possible causes. Finally, evaluate the effects of misconfigurations on the reachability of a zone's network. Our results show that, although progress has been made in the implementation of DNSSEC, over 4% of evaluated domains show misconfigurations. Of these misconfigured domains, almost 75% were unreachable from a DNSSEC aware resolver. This illustrates that although the authorities of a domain may think their DNS is secured, it is in fact not. Worse still, misconfigured domains were at risk of being unreachable from the clients who care about and implement DNSSEC verification while the publisher may remain unaware of the error and its consequences.

2.3 CACHE FUNCTION ACTIVATION ON A CLIENT BASED DNSSEC VALIDATION AND ALERT SYSTEM BY MULTI THREADING

Kakoi, K., Jin, Y., Yamai, N., Kitagawa, N., & Tomoishi, M. proposed the model Domain Name System (DNS) was one of the most important services of the Internet since most communications normally begin with domain name resolutions provided by DNS. However, DNS has vulnerability against some kind of attacks such as DNS spoofing, DNS cache poisoning, and so on. DNSSEC was an security extension of DNS to provide secure name resolution services by using digital signature based on public key cryptography. However, there were several problems with DNSSEC such as failing resolution in case of validation failure, increasing the load of DNS full resolver, and so on. To mitigate these problems, proposed a Client Based DNSSEC

Validation System. This system performs DNSSEC validation on the client, and in case of validation failure, it forwards the failed response and alerts the user to the fact. However, this system has a problem that it inactivates the cache function of validation library so that it always performs DNSSEC validation even for the same query. In this paper, reported how to solve this problem by multithreading of DNSSEC validation system.

2.4 AN ADVANCED CLIENT BASED DNSSEC VALIDATION AND PRELIMINARY EVALUATIONS TOWARD REALIZATION

Jin, Y., Tomoishi, M., & Yamai, N. in this paper DNSSEC (Domain Name System Security Extensions) was designed to provide security functions for the current DNS protocol. However, DNSSEC yet has low deployment rate in the Internet due to its heavy workload on DNS full resolvers and high administrative cost. Furthermore, DNSSEC does not cover the last one mile in name resolution: between the DNS full resolver and client. In order to provide complete DNSSEC service between authoritative zone servers and clients, a new DNSSEC validation mechanism with acceptable workload on DNS full resolver and client was required. In this paper, propose an advanced client based DNSSEC validation mechanism and compare the DNSSEC performance between DNS full resolver and client based on evaluations in a local experimental network. By validating DNSSEC on each client, the proposed mechanism can reduce the workload of DNS full resolvers and also can provide secure name resolution for each client. According to the results of preliminary evaluations and confirmed that it is possible to reduce the workload of DNS full resolver by transferring the DNSSEC validation process to clients with acceptable extra workload. More importantly, the benefit of DNSSEC can be extended to clients with secure name resolution service.

2.5 PROBLEM FORMULATION

In ECC, the security is achieved only if cryptographically strong elliptic curves are used and there is a lack of protection between user devices and resolvers. DNSSEC protects only Resource Records if they are authoritative in the zone. DNSSEC increases computational requirements on validation and Servers. It does not provide end to end security. DNSSEC suffers from availability problems due to packet fragmentation and is a potent source of distributed denial of source attacks.

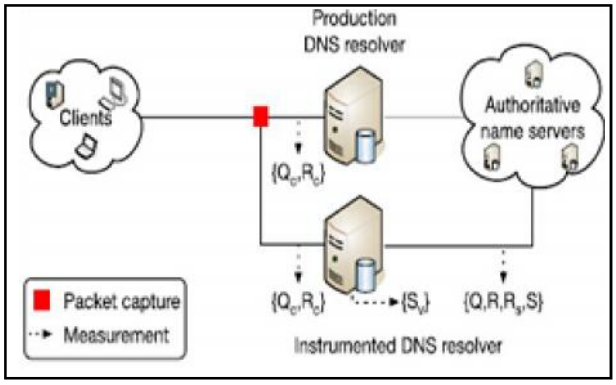
In ECC signature validation much slower than the RSA validation. Simple linear regression increases the complexity to measure the each parameter. The load balancing is the complexity issue in linear regression model and this leads to reduce the stability of the network.

3. EXISTING WORK

In Existing System, analyze the switching from DNSSEC to ECC lead to handle problems for DNS resolvers. So, utilized the new model that accurately predicts how many signature validations DNS resolvers have to perform. This allows us to calculate the additional CPU load ECC imposes on a resolver. Using real-world measurements from four DNS resolvers and with two open-source DNS implementations, then evaluate future scenarios where DNSSEC is universally deployed. The performance of the new model evaluated through performance parameter and simple linear regression (SLR) used to measure the each parameter. The results shows conclusively show that switching DNSSEC to ECC signature schemes does not impose an insurmountable load on DNS resolvers, even in worst case scenarios.

3.1 Resolver Implementation

The model works for different DNS resolver implementations, then compared two popular open source packages. The first is Unbound, developed by NLnet Labs. Unbound is a resolver only implementation, designed from the ground up to support DNSSEC validation, and optimised for speed. The second is BIND, the oldest and most popular open source DNS implementation. BIND implements both resolver and authoritative name server functionality in a single application.



Based on the measurement setup, two instrumented resolvers were deployed, one running Unbound, the other running BIND. Both resolvers are run simultaneously were fed live client data from production resolver. Based on the parameters r and s are almost identical for the two resolver implementations. That the both resolver sent the same query stream this is expected. Second, shows a difference in the fraction of responses that contain signatures (αs).This is due to implementation differences between Unbound and BIND. Third, as shows, the most significant implementation difference between Unbound and BIND immediately becomes apparent when we perform the parameter estimation. The main takeaway is that the model works for the two different resolver implementations.

3.2 Stability Over Time

Predictions are only meaningful if the parameters of the model remain stable over time. In particular, r, s and αv should not change much over time. Based on the parameter estimation through linear regression time t1 ,t2 and t3 data was captured over a period. The only noticeable fluctuations occur for s and αv at t2. This fluctuation is self-cancelling, because if s rises αv while decreases the net effect on a prediction for the total model is negligible. This fluctuation is self-cancelling, because if *s* rises while decreases the net effect on a prediction for the total model is negligible.

3.3 Different Client Population

To evaluate how well the model works for differing client populations, we performed parameter estimations based on measurements for all four resolvers *r*1...*r*4 described in below table.

URL	$\bar{r}$	α_v	α_s	$\bar{s}$
Kongu.ac.in	18.0	16.0	3.0	5.333
Facebook.com	17.0	15.0	3.0	5.0
Twitter.com	15.0	15.0	2.0	7.5
Google.com	15.0	12.0	2.0	6.0
Amazon.com	16.0	14.0	2.0	7.0
Microsoft.com	18.0	11.0	2.0	11.0
Yahoo.com	15.0	13.0	1.0	13.0

Despite having different client populations of different sizes, as can already seen, the parameter estimations for r1…r3 lead to almost the same values for r, s and αv. The only variation is observed for αs which, as mentioned above, does not influence the predictive capabilities of the model. While we see few differences between r1…r 3, there is a noticeable difference between these three resolvers and r4 . Figure shows a comparison between the parameter estimation for r1 and r4 . As the figure shows, only r is roughly the same, while the other two important parameters, s and αv differ significantly. There are two explanations for this. First, the query name popularity for differs r4 from that for r1;

just like the difference between times t_1 ,t_2 and t_3, this most likely means that r_4 receives more responses with signatures in the additional and authority sections. Second, and more importantly, the client population and query load for r_4 are much smaller than for the other three resolvers. This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination. This is reflected in the scatter plots for measurement results in which show that the blue scatter points for r_4 are bunched much more tightly together towards the bottom left of each of the four subplots. This then, is a shortcoming of the model: it will tend to be less accurate for DNS resolvers with a lower query load.

3.4 Predictive Qualities

Compute coefficient of determination

$R^2 = 1- \Sigma i\ (yi – fi)^2/ \Sigma i\ (yi - yi)^2 \quad 0<R^2<1$

Here R^2 is a measure for the fraction of the variance in the observed data. In general a higher value for R^2, closer to 1, indicates a better fit, and thus a better model. Then, using the observed value for *Q* (the number of outgoing queries from the resolver), we used the model to predict how many signatures would need to be validated (S_v predicted). Then compared this to the number of signatures that were actually validated (S_v observed) and computed R^2. Finally conclude that the model is a good predictor of the number of signature validations (S_v) that need to be performed given a certain number of outgoing queries (*Q*) from a DNS resolver. That the DNS resolver to which the model is applied must have a sufficiently large client population and a sufficiently high query load. Given that a large client population and high query load constitute a worst-case scenario in terms of the expected number of signature validations, this makes the model well-suited to analyse the impact of ECC signature validation onvalidating DNS resolvers.

4. ALGORITHM

4.1 Simple Linear Regression Algorithm

A regression line is a straight line that attempts to predict the relationship between two points, also known as a trend line or line of best fit. Simple linear regression is a prediction when a variable (y) is dependent on a second variable (x) based on the regression equation of a given set of data. While various non-linear forms may be used, simple linear regression models are the most common.

SUB Regress (x,y,n,a_1,a_0,syx,r_2)
Sumx = 0:sumxy = 0:st = 0
Sumy = 0:sumx$_2$= 0:sr = 0

DOFOR i = 1,n

sumx = sumx+xi
sumy = sumy+yi
sumxy = sumxy+xi*yi
sumx2 = sumx2+xi*xi

END DO

xm = sumx/n
ym = sumy/n
a_1= (n*sumxy-sumx*sumy)/(n*sumx2-sumx*sumx)
a_0 = ym-a1* xm

DOFOR i=1,n

st = st + $(y_i-y_m)^2$-^{2}
sr = sr + $(y_i-a_1*x_i-a_0)^2$

END DO

syx = $(sr/(n-2))^{0.5}$
r2= (st - sr)/st

END Regress

x and y are the variables.
a_1 = The slope of the regression line
a_0 = The intercept point of the regression line and the y axis.
n = Number of values or elements
X = First Score
Y = Second Score

sumxy = Sum of the product of first and Second Scores
sumx = Sum of First Scores
sumy = Sum of Second Scores
sumx2 = Sum of square First Scores

Given that the plots suggest linear relationships between the variables, the model for validation resolver is defined a set of parameterized linear functions $f_1 \ldots f_4$ specified below:

The set of parameterized linear function f1,f2,f3,f4
f1: $R=rQ+\beta_1$ $_1$ f3 : $S=sR_s+\beta_3$ $_3$
f2: $Rs=\alpha sR+\beta_2$ $_2$ f4 : $Sv=\alpha vS+\beta_4$ $_4$ with:

r- the average number of responses per query αs- the fraction of responses with signatures

s - the average number of signatures per response αv - the fraction of signatures that is validated These functions can then be combined to give *f: f* : $S_v = aQ + b$

$= \alpha_v s\alpha_s r$
$= \alpha_v (S(\alpha_s\beta_1+\beta_2)+\beta_3)+\beta_4$
Linear Regression: Y_i $_0$ $_1$ X_i $_i$

Y_i - Outcome of Dependent Variable (response) for i^{th} experimental/sampling unit
X_i - Level of the Independent (predictor) variable for i^{th} experimental/sampling unit
0 1 X_i - Linear (systematic) relation between Y_i and X_i (aka conditional mean)
0 - Mean of Y when X=0 (Y-intercept)
1 - Change in mean of Y when X increases by 1 (slope)
i - Random error term

Note that 0 and 1 are unknown parameters. We estimate them by the least squares method.

Obtain a sample of *n* pairs $(X_1,Y_1)\ldots(X_n,Y_n)$.
Plot the Y values on the vertical (up/down) axis versus their corresponding X values on the horizontal (left/right) axis.

5. PROPOSED WORK

The number of outgoing queries from a resolver is one among the most determinants of the number of signature validations that a validating DNS resolver needs to perform. The outgoing question rate may be a operate of the number of queries from client and query name quality. Queries from client can solely result in outgoing queries from the resolver if the solution isn't already cached. Thus, although popular domains is also queried a lot of times by client, this doesn't essentially result in a high outgoing question rate. One development which will modification this can be the big scale introduction of recent generic top-level domains (gTLDs)[32]. If these new gTLDs prove to be popular, this might result in a larger spread in names on the net, which can reduce the effectiveness of caching by resolvers and result in higher numbers of outgoing queries. This could be studied in future work, as a bigger variety of outgoing queries can result in a higher variety of signature validations. V RESULTS AND DISCUSSION The results shows conclusively show that switching DNSSEC to ECC signature schemes does not impose an insurmountable load on DNS resolvers, even in worst case scenarios. There are two scenarios for DNSSEC deployment 1.Current DNSSEC deployment 2.Popular-domains-first to 100% deployment

6. CURRENT DNSSEC DEPLOYMENT

In this scenario, the peak signature validation rate observed on the busiest resolver. The highest rates measured were observed in the measurement at 5ШаЬ2. For the Unbound resolver implementation, validation peaked at 124 signatures per second, for BIND it peaked at 224 signatures per second. The maximum signature validation rates that can be achieved with each of the benchmarked ECC signature schemes. In other words, if all of the current DNSSEC deployments on the Internet were to switch to an ECC-based signature scheme overnight, this would not pose a problem for validating DNS resolvers, and would leave ample room for growth both in terms of DNSSEC deployment as well as an increase in query load on the resolver

7. POPULAR-DOMAINS-FIRST TO 100% DEPLOYMENT

For this evaluation, measured query name popularity for outgoing queries from a DNS resolver. The reason that the query name popularity on the outgoing side was chosen is that this represents the absolute worst-case scenario for the resolver for which the distribution is measured. On the outgoing side, popularity is not just determined by popularity of the name among the client population of the resolver, but is also determined by the time-to-live (TTL) of records for certain names. Moderate popularity on the client side combined with a low TTL for DNS records will lead to a high number of outgoing queries (to refresh the cache).

8. CONCLUSION

In this paper we've got once and for all answered the question can validating DNS resolvers handle the extra mainframe load imposed by the validation of elliptic curve-based signatures. We show that a collection of linear relationships accurately models the behaviour of a validating DNS resolver. By this model, we are able to

dependably predict future developments in signature validation. By combining these results with benchmarks of various elliptic curve digital signature schemes we have got shown that the mainframe needs for signature validations don't exceed the capability of one serious concern core, even though the most CPU-intensive elliptic curve is used. We mentioned remaining hurdles that operators desire to switch to ECC-based signature schemes might encounter, such as support for elliptic curve cryptography keys by TLD registries and name registrars. we tend to believe these issues to be transient; all are in the process of being resolved by the net community. We additionally mentioned another serious concern, raised by an operator, that is that the potential for denial-of-service on a validating DNS resolver through validating starvation. This threat requires the implementers of validating DNS resolver package, and can be able to implement effective countermeasures by applying some type of rate limiting. The use of ECDSA scheme in DNSSEC package has important advantage and needs to support ECC signature schemes. All imagine implementations support ECDSA P-256 and P-384. Support for the newer algorithms presently being standardized (Ed25519 and Ed448), however, is almost non-existent. Operators may have to upgrade to newer versions of DNSSEC signer package to achieve elliptic curve cryptography support.

9. FUTURE WORK

In existing work the security is achieved only if cryptographically strong elliptic curves are used and there is a lack of protection between user devices and resolvers. DNSSEC protects only Resource Records if they are authoritative in the zone. DNSSEC increases computational requirements on validators and Servers. It does not provide end to end security. DNSSEC suffers from availability problems due to packet fragmentation and is a potent source of distributed denial of source attacks.ECC signature validation much slower than the RSA validation. The disadvantage is Simple linear regression increases the complexity to measure the each parameter. The issue can be overcome with using various algorithms and it reduce the burden of authoritative name servers to validate the unnecessary queries. It increase the security and Perform well in large network.

REFERENCES

[1] G. van den Broek, R. M. van Rijswijk, A. Sperotto and A. Pras, "IDNSSEC Meets Real World: Dealing With Unreachability Caused By Fragmentation", IEEE Commun. Mag., Vol.52, No.4, Apr. 2014, pp.154-160.

[2] W. Lian, E. Rescorla, H. Shacham and S. Savage, "Measuring the Practical Impact of DNSSEC Deployment", in Proc. 22nd USENIX Secur.Symp. (USENIX Security), Washington, DC, USA, 2013, pp. 573–588.

[3] R. van Rijswijk-Deij, A. Sperotto, and A. Pras, "IMaking the Case for Elliptic Curves in DNSSEC", ACM Comput. Commun. Rev., Vol.45, No.5, Oct. 2015, pp.13-19.

[4] R. van Rijswijk-Deij, M. Jonker and A. Sperotto, "On the Adoption of the Elliptic Curve Digital Signature Algorithm (ECDSA) in DNSSEC", in Proc. 12th Int. Conf. Netw. Service Manage. (CNSM), Montréal, QC, Canada, 2016.

[5] D. Wessels, M. Fomenkov, N. Brownlee and K. Claffy, "Measurements and Laboratory Simulations of the Upper DNS Hierarchy", in Proc. 5thInt. Workshop Passive Active Meas., 2004, pp. 147-157.

[6] Y. Koç, A. Jamakovic, and B. Gijsen, "A Global Reference Model of the Domain Name System", Int. J. Critical Infrastruct. Protection, Vol.5, No.3-4, Dec. 2012, pp. 108-117.

[7] H. Yang, E. Osterweil, D. Massey, S. Lu and L. Zhang, "Deploying Cryptography in Internet-scale Systems: A Case Study on DNSSEC", IEEE Trans. Depend. Sec. Comput., Vol. 8, No.5, Sep./Oct. 2011, pp. 656-669.

[8] B. Laurie, G. Sisson, R. Arends and D. Blacka, "DNS Security (DNSSEC)Hashed Authenticated Denial of Existence", Document RFC 5155, 2008.

[9] D. York, O. Surý, P. Wouters and O. Gudmundsson, "Observations on Deploying new DNSSEC Cryptographic Algorithms", Tech. Rep., 2016.

[10] T. Halvorson *et al.*, "From Academy to Zone: An Analysis of the new TLD Land Rush", in Proc. ACM IMC, Tokyo, Japan, 2015, pp. 381-394.

Improving Networks Lifetime Using PSO Algorithm in WSN

C.Visali and J.Premalatha
Department of Information Technology, Kondu Engineering College, Perundurai, Erode- 638 060, Tamil Nadu
E-mail: visalisugu@gmail.com, jprem@kongu.edu

Abstract

In physical buildings, for example, water/manage diffusing networks, are obvious by way of battery-managed wireless Sensor Networks (WSNs). Due to the fact battery substitution of sensor focuses is on the whole troublesome, whole association gazing may also be just subtle if the operation of the WSN focuses adds to a long WSN lifetime. Two discernible tactics to lengthy WSN lifetime are I) consummate sensor authorizing and ii) suitable knowledge collecting and sending in context of compressive perceiving. These approaches are possible simply if the authorized sensor focuses hooked up a related correspondence maintain (connectivity constraints), and fulfill a compressive recognizing interpreting important (cardinality predominant). These two necessities make the obstacle of extending structure lifetime via sensor focus factor start and compressive recognizing NP-rough. Clustering sensor hubs is a powerful topology manipulate procedure conducting this goal. On this paper, we display another approach to pull out the procedure lifetime in mild of the improved molecule swarm development calculation, which is an growth system meant to decide upon goal hubs. The convention considers both vitality effectiveness and transmission separation, and hand-off hubs are utilized to mitigate the exorbitant vigor utilization of the group heads. The proposed conference brings about better dispersed sensors and an all-round adjusted clustering framework upgrading the process's lifetime. We distinction the proposed convention and similar conventions through shifting more than a few parameters, e.g., the range of hubs, the system territory measure, and the function of the bottom station. The examination uncovers that the two streamlining problems provide precise systems, but the many-sided quality between the lifetime accomplished by using the vitality editing strategy and the satisfactory lifetime is little when the most important essentialness at sensor focal point focuses is in a basic experience extra prominent than the importance utilized for a solitary transmission. The lifetime satisfied through the essentialness altering is asymptotically immaculate, and that the potential system lifetime isn't any no longer as much as 1/2 of the best. Examination and numerical redirections evaluate the gainfulness of the proposed vitality altering system.

Keywords: *Connectivity constrains, Cardinality constraints, Clustering techniques, Wireless sensor networks*

1. INTRODUCTION

Wireless Sensor Networks (WSNs) are being used to monitor giant buildings in sharp urban social events, for illustration, sections, and towers [1]. When you consider that sensor middle concentrations are on the whole control confined, and battery substitution is difficult or even perpetual, design lifetime is a primary Execution metric. Just a few methods have been proposed to drag out structure lifetime and alongside these lines to interact entire course of action checking.

For instance, sensor middle concentrations can shape get-togethers, the place sensor center concentrations can graph social affairs, the place sharing attention publications take swing toward going about as percent go to alter the centrality usage of within focuses [2], [3], [4]. The core focuses can revive organizing [5], [6] or utilize multi-soar brief range correspondence [7] to spare hugeness transmission. Simulation results [8] can be utilized to slash the transmitted information volume.

The ways to be utilized for centrality sparing have to rely upon the attributes of the checking applications. Thick sensor build has the going for walks with principal fixations: 1) higher zone of occasions; 2) structure three) lessened basics use in expertise transmission via manhandling multi-ricochet brief variety correspondence. On this method, regardless of the manner that thick methods exhibit a better institution fee, they generously slash the preservation rate in sort, and on an awfully basic stage, may give better-checking execution. Power usage can be correctly overseen via enhancing the gadget topology and controlling the hubs' transmission manage degrees within the directing convention [15], [16].

The clustering strategies procedure is precious in diminishing pressure usage in directing conventions [17]. In a clustering strategies layout, sensor hubs are composed into organizations, wherein the sensor hubs with bring down energy can be applied to perform detecting errands, and send the detected statistics to their bunch head at a quick separation [18].

A hub in a set may be picked as the bunch head (CH) to wipe out associated statistics from the people from the group, with the goal of diminishing the measure of the totaled information transmitted to the BS [18], [16].

The clustering strategies method can build organize existence span and to enhance vitality productivity by restricting standard power utilization and adjusting vitality utilization a number of the hubs amid the device lifetime [16], [18]. in addition, it is equipped for easing channel dispute and package crashes, bringing approximately better system throughput under high load [18], [12].

2. LITERATURE SURVEY

2.1 Lifetime Maximization by Flow of Data

The lifetime of a framework phenomenally depends upon the waiting essentialness of the sharing center features. There are specific items for the noteworthiness use of sensor focal point focuses. In the centrality utilize is straightly recognized with the tolerant power, transmitting force and know-how transmission expense, and the customary lifetime of a sensor center of attention factor is depicted because the degree of the vitality uttermost scopes of the inside and the run of the mill essentialness use. In this mannequin, the vitality utilization of the focuses depends likewise on the parcel (hop rely).

In [9], the significance makes use of the middle focuses in relaxation mode are permitted to be zero, we arrange the vitality utilization of the dynamic sensor center of attention focuses to 1 and set that of the within focuses in relaxation mode proportionate .Within the value utilize is straightly identified with the tolerant power, transmitting drive and knowledge transmission rate, and the traditional lifetime of a sensor focus is depicted as the degree of the hugeness farthest reaches of the middle point and the general centrality utilization.

In [9], the essentialness usages of the centers in leisure mode are believed to be 0, we institutionalize the imperativeness use of the dynamic sensor facilities to 1 and set that of the middle points in leisure mode equal to zero. Other than the nodal point of view, drawing out lifetime from the framework perspective, e.g., with the aid of redesigned coordinating, has before probably the most phase suggestion about. In this great trouble, the enormous movement estimation is conventionally used [15], [16].

In the first work [15], the essentialness use of the framework has been exhibited as a factor of the movement stream directing selections. By using then the difficulty is given an element as a straight programming hindrance.

In a an identical framework environment, the place every sensor core point can both transmit its information to its neighbor with low imperativeness price, or transmit knowledge direct to the sink core with excessive essentialness cost, boosting framework lifetime is indistinguishable to flow extension and imperativeness enhancing [16]. In such circumstance, essentialness changing has been used to develop mastermind lifetime [17], [18].

Other strategy to take care of alter the imperativeness use is turning the working time of sensor middle elements, i.e., empowering some sensor facilities to relaxation without surrendering within the checking execution have regarded Discovering unmistakable related preparations of the WSN to drag out framework lifetime.

In every timeslot, conveniently the sensor centers within the related administering set are dynamic and interchange core features are put into relaxation.

Christos G. Cassandras (2014) proposed a gold standard manipulate technique is used to remedy the concern of routing in sensor networks where the goal is to maximize the network's lifetime. The vigor sources (batteries) at nodes are usually not assumed to be "superb" however alternatively behaving according to a dynamic vigour consumption mannequin, which captures the nonlinear conduct of specific batteries.

In a fixed topology case there exists a most efficient coverage together with time-invariant routing probabilities, which is also bought by means of solving a collection of reasonably simple nonlinear programming (NLP) issues. This most advantageous policy is beneath very slight conditions, powerful with appreciate to the battery mannequin used.

The technique to this hindrance is given with the aid of a coverage that depletes all node energies whilst and that the corresponding power allocation and routing probabilities are received with the aid of solving an NLP trouble.

2.2 Compressive Sensing for Data Gathering

Considering that a thick framework, in our earlier work [14], we proposed the information gathering design showed up in Fig. 1, where just about the stupid attention centers are dynamic and transmit knowledge within the CDG path to the sink core factor.

Due to the fact that each and every robust attention point transmits a readied vector in surroundings of the summation of its estimation and its acquired vector, the bundle sizes of the concentrations are the equal, and the essentialness use of the dynamic attention facilities is balanced.

If the endorsement of the sensor centers is picked unequivocally in each and every checking timeslot, the significance use of all the sensor bases will also be on balanced.

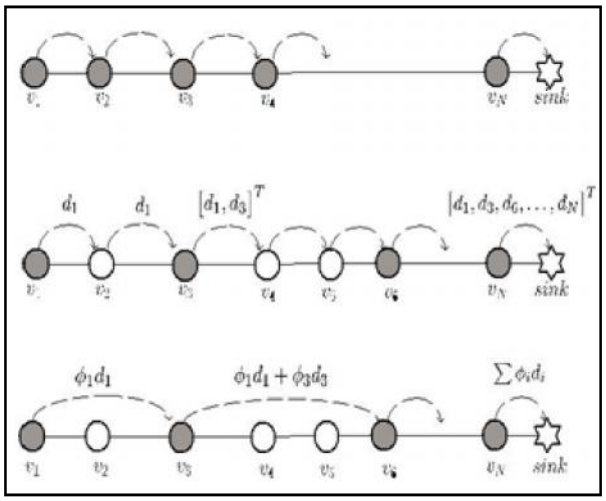


Fig.1 Compressive Sensing

L. Xiang, J. Luo, and A. Vasilakos (2011) compressed sensing (CS) is being more and more applied to wireless sensor networks communications. Utility of CS to information assortment in wireless sensor networks, and aim at minimizing the community energy consumption through joint routing and compressed aggregation. Represent the top-rated method to this optimization obstacle, then we prove its NP-completeness. Proposed a mixed-integer programming system along with a grasping heuristic, from which each the most efficient (for small scale problems) and the close-foremost (for giant scale issues) aggregation trees are acquired.

2.3 Clusteing Techniques

Bahbahani, M., and Alsusa, E. (2017), proposed a valuable grouping convention in light of the low vitality versatile bunching development technique to take care of reinforce the existence span of vitality reaping established far off sensor techniques (EH-WSN). In the proposed conference, to warranty that any vitality utilization related with the a part of the staff head (CH) is shared between the hubs, the CH phase used to be substituted between the hubs utilizing responsibility cycling as an element of their individual vitality Collecting capacities.

Additionally, to maintain up a vitality nonpartisan operation when now not going about as a CH, the hubs obtain an know-how transmission duty cycle and any overabundance vitality was once put assets into handing-off other hubs' bundles. To enhance the handing-off execution, a novel cross-layer important TDMA conspire used to be likewise presented. The best quantity of organizations in an EH-WSN is broke down as far as vitality utilization, inactiveness, and transmission capacity use. Reproductions, carried out utilizing green Castalia, exhibit unmistakable execution enhancements in embracing the proposed conference over benchmark conspires so far as throughput and lifelong, certainly below profoundly compelled vitality stipulations.

Hong, Z., Wang, R., & Li, X.(2015), awarded easy methods to plan a vitality productive calculation to develop the process lifetime in entangled occasions is a normal hassle for heterogeneous remote sensor systems (HWSN). A bunching tree topology manage calculation in view of the vitality conjecture (CTEF) used to be proposed for sparing vitality and guaranteeing system stack adjusting, at the same time while desirous about the connection first-class, bundle misfortune fee, and many others. In CTEF, the common vitality of the procedure was exactly expected per round (the lifetime of the system used to be signified via rounds) as far as the contrast between the best and actual common lingering vitality making use of focal factor of confinement hypothesis and natural circulation instrument, the entire while. On this premise, team heads had been chosen by price work (counting the vitality, interface nice and bundle misfortune cost) and their separation. The non-bunch heads have been resolved to join the staff via the vitality,

separation and connection first-class. Moreover, a number of non-bunch heads in every group had been picked as the switch hubs for transmitting knowledge via multi-jump correspondence to decrease the heap of each bunch head and drag out the lifetime of the process. The reproduction comes about show the skill ability of CTEF.

2.4 Problem Formulation

We recall a WSN including two ranges of focuses that screens a sector of line form. Trademark such outlines are a pipeline in a water dispersal shape, a place, or a system.

The fundamental degree contains battery-managed sensor recognition focuses which can be thickly exceeded on in the checked area.

The second level carries sink cognizance focuses, which can be arrange managed and are sent at the 2 fruitions of the line. They gather facts from the sensor focuses, and transmit the data to a far flung watching recognition. In angle of the length of the checked vicinity and the close by little correspondence degree of the sensor attention factor a multi-bypass correspondence course from the sensor focuses to the sink focus must be set up.

Since battery substitution isn't fundamental for the applications said more than, a basic target is to amplify the structure lifetime. Naturally, it is noteworthy to keep alive however an incredible piece of the sensor focus focuses as could be ordinary, which drives the system of organization calculations in light of criticalness adjusting, i.e., ideally actuate the focuses with all the all the more extraordinary vitality.

Hence, the basic issue to be considered here, is whether the most over the top structure lifetime can be master by the vitality altering approach, or (if not), what is the execution of the importance changing philosophy as for system lifetime

3. EXISTING WORK

3.1 Coneectivity Constraints

We characterize the lifetime of a WSN to be the working time until either WSN winds up noticeably separated, or the observing execution of the WSN can't be ensured. In each timeslot, the availability and the observing execution necessity of the dynamic sensor organize must be fulfilled. Let twofold factor xi(t) show whether hub vi is dynamic at timeslot t.

At that point, the vitality elements of vi can be composed as Ei(t + 1) = Ei(t) xi(t) and the scheduling problem considered in this paper is to determine x(t) = [x1(t),...,xN(t)]T,"Give G(x(t)) a chance to mean the prompted chart of dynamic sensor hubs and the sink hubs.

Definition 1: (Availability Imperative) The actuation of the sensor hubs x(t) fulfills the network limitation if and just if the prompted chart G(x(t)) is associated.
Definition 2: (Cardinality Imperative) The initiation of the sensor hubs x(t) fulfills the cardinality limitation if P furthermore, just if xi(t) Mcs, where Mcs is controlled by the required estimation mistake of the deliberate information.

At that point, the lifetime augmentation issue can be defined as an ideal control issue as takes after:
Max " xi p
x i"v
s.t. " xi=max {Mcs,Mc} G(x) is connected,
Xi " {0, 1}, "i "V

Definition 3: (Initiation Profile) A sanctioning profile is a social affair of sensor center points that satisfies the accessibility necessity. We say an institution profile is feasible if and just if it in like manner satisfies the cardinality necessity.

3.2 Energy Balancing Problem

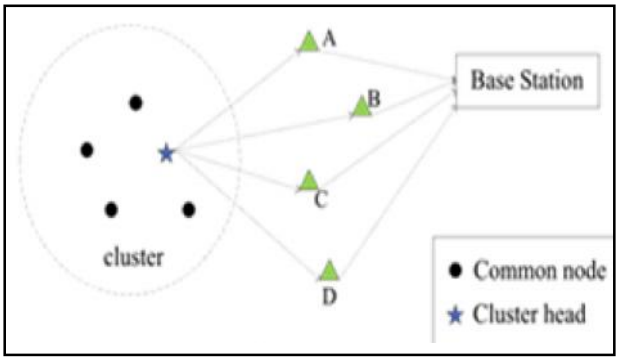
In our past work [14], we proposed a vitality adjusting issue, together with an answer technique. Since in this paper we examine the crucial properties of the vitality adjusting issue from the perspective of system lifetime expansion, we give the essential points of interest in the accompanying .

pi(t) = Ei(t)=Ei In [14], we built up a calculation to take care of Issue (4). The points of interest of the methodology are appeared in Calculation 1. To begin with, Calculation 1 discovers Mc by a briefest way calculation, for example, Dijkstra's calculation in Line 1, to be specific finds the most limited way from v0 to vN+1, where the weights of the considerable number of edges are 1. At that point, the base number of sensor hubs, m, that fulfills both the availability and the cardinality requirements is computed in Line 3.

ALGORITHM 1
INPUT : Adjacency matrix A, the minimum number of active node Mcs the normalized residual energy of the sensor nodes p.
OUTPUT : A set of sensor nodes VA that need to be activated.
PROCEDURE
1. Find the minimum number of sensor nodes Mc ,that satisfy the connectivity constrains
2. If Mc<+6 then
// Find the minimum number of nodes that satisfy both connectivity and cardinality constraint
3. m = max{Mc, Mcs}
4. Calculate g(s1,m)
5. Return №”a
6. Else
Return I”
7. End if
ALGORITHM 2
INPUT : Adjacency matrix A, the minimum number of active node battery of the nodes Mcs, the battery of the nodes Ei, “i.
OUTPUT : Network lifetime
PROCEDURE
1. Set t ђ! 1, Flag ђ! TRUE, Ei(t) = Ei
2. while Flag do
3. Set pi ђ! Ei(t)/Ei, p = {p1,...,pn}
4. Find Mc for the connectivity constraint
5. VA ђ! Call Algorithm 1 with input A,Mcs,p
6. if VA ‘“ “ then
7. Set t ђ! t + 1, Ej(t) ђ! Ej(t”1)”1,”j “V
8. Else
9. Set Flag ђ! FALSE
10. end if
11. end while
12. return T ђ! t”1

4. PROPOSED WORK

In our convention, hubs are arranged into the CHs, transfer hubs (RNs) and basic hubs (CNs). The operation of the convention incorporates two stages, i.e., the grouping setup stage and information transmission stage. The two stages are performed in each round of the system operation and rehashed intermittently. In the bunching setup stage, the groups, CHs and RNs and additionally the way between each bunch and the sink (or the BS) are resolved, and afterward the system is sorted out. In the information transmission stage, the CHs gather information from all the group individuals and exchange to the transfer nodes which then relay the data to the BS according to the topology determined in the phase with the coverage range.



Relay nodes’ selection

- At the beginning, each node sends a Node-MSG message to broadcast its residual energy information what’s more, area data, which are fundamental for choosing the group heads and hand-off hubs; The BS chooses the bunch heads by utilizing the calculation and communicate a message including the group heads’ ID to educate the system of the group head’s area. After the group heads know their status, each bunch head acquaints itself with the system by communicating a little commercial message (i.e., CH-ADV), which utilizes the non-tireless bearer sense various access (CSMA) MAC convention. The message incorporates the group head’s ID and a header that distinguishes it as an ad message; Then, also, the BS select the hand-off hub by utilizing the calculation Once a transfer hub is chosen, a notice message (i.e., RN-ADV), which incorporates its ID, the comparing group head’s ID and the header, is sent to the system by the BS to announce its status as a hand-off hub. Every basic hub chooses its group by picking the bunch head that requires the base transmission vitality, in light of the quality of the CH-ADV message from each bunch head. At that point, a group is picked; after every normal hub has chosen which bunch it participate, it must educate the group leader of its choice by transmitting a JOIN-REQ message. The message is again short, comprising of the hub’s ID, the having a place group head’s ID and the sender’s leftover vitality

- The bunch head in a group goes about as the control place for the goal of organizing information transmissions. The bunch head sets up a TDMA scheduler and communicates the SCHEDULE-MSG message to the regular hubs in the group and in addition the relating transfer hub. This maintains a strategic distance from impacts among information

messages, and furthermore permits the radio segments of every normal hub and transfer hub to be turned off constantly, with the exception of when the regular hubs transmit messages or hand-off hubs get messages. This causes us to increment phantom proficiency and diminishing vitality utilization by singular sensors.

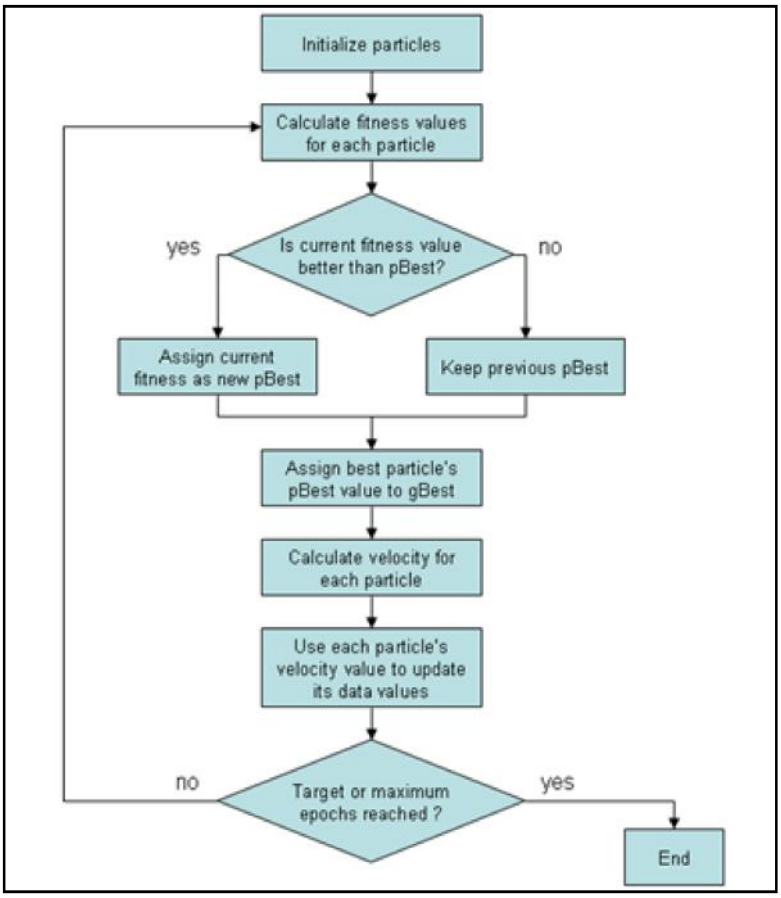
4.1 Improved Particle Swarm Optimization Technique

Owing to its simple idea and high efficiency, PSO has grow to be a widely followed optimization approach and has been efficiently implemented to many actual-international troubles, specifically multimodal problems [12], [13]. Hence, it's far an powerful algorithm to solve the clustering troubles of electricity efficiency and minimal transmission distances for the clustering setup section.

In our preceding work, we use PSO set of rules to resolve software program-described community troubles efficaciously [40]. However, PSO plays poorly in phrases of nearby search with premature convergence, particularly for complicated multi-peak search problems [41], [42]. In order to deal with this particular situation, we improved the traditional PSO algorithm by using adjusting the inertial weight to avoid particles being trapped in neighborhood optima, and used the stepped forward PSO set of rules to maximize the fitness capabilities of (five) and (eight). The technique consists of the following 5 fundamental steps:

- Initialize the optimization problem and algorithm parameters.
- Calculate the fitness values
- Update velocity and position vectors
- Change the inertial weight.
- Go to step 3 until the termination criterion is met

Parameters Node	Existing 70	Distance 150m	Proposed 100	Distance 230m
node	Scenario 1	Scenario 2	Scenario 1	Scenario 2
50	20	30	40	50
100	30	30	50	60



5. RESULTS AND DISCUSSION

The performance of Qos is analyzed by implementing connectivity and cardinality constraints with energy balancing. The result shows that Qos performance is improved 60% The following parameters taken into account for Qos performance are

- Throughput
- Delivery ratio
- Energy

6. CONCLUSION

In work proposed an imperativeness modifying methodology in light of perfect authorization logbook and start in perspective of essentialness altering figuring and compressive distinguishing. Authorizing particular sensor favorable circumstances to the framework field. It lessens the imperativeness. By then the coordinating method used is clear and fiery. An institution of particular sensor system is exhibited which reduces the imperativeness. These segments achieved an ok execution in the framework. Some of parameters are used for propagation.

They are throughput, package transport extent, and total residual imperativeness and so forth. These parameters are evaluated with different regards for the execution examination.

In existing technique there is an essentialness altering strategy in light of the figuring and compressive identifying. It uses the gainful imperativeness when outline with the other computation .Still there is an open issues in essentialness altering.

7. FUTURE WORK

In existing method there is a centrality adjusting system in light of the check and compressive recognizing. It utilizes the profitable noteworthiness when chart with the other estimation. Still there is an open issues in vitality evolving.

The issue can be overcome with using various algorithms by applying in hybrid clustering techniques and hierarchal compressive sensing to improve more lifetime by balancing energy.

REFERENCES

[1] H.S. AbdelSalam and S. Olariu, “Toward Adaptive Sleep Schedules for Balancing Energy Consumption in Wireless Sensor Networks”, IEEE Transactions Computers, Vol.61, No.10, 2012, pp.1443-1458.

[2] J. H. Chang and L. Tassiulas, “Maximum Lifetime Routing In Wireless Sensor Networks”, IEEE Transactions Networking, Vol.12, No.4, 2004, pp.609-619.

[3] G. Degirmenci, J.P. Kharoufeh and O. Prokopyev, “Maximizing the Lifetime of Query-Based Wireless Sensor Networks”, ” ACM Trans. Sensor Networks (TOSN), Vol.10, No.4, 2014, pp.56:1-56:24.

[4] R. Du, L. Gkatzikis, C. Fischione and M. Xiao, “Energy Efьicient Sensor Activation for Water Distribution Network Based on Compressive Sensing”, IEEE Journal on Selected Areas in Communications, Vol.33, No.12, 2015, pp. 2997-3010.

[5] Jiao Zhang and Tao He, “EBRP: Energy-Balanced Routing Protocol for Data Gathering in Wireless Sensor Networks”, IEEE Transactions on Parallel and Distributed Systems, Vol.22, No.12, 2015.

[6] C. Karakus, A.C. Gurbuz and B. Tavli, “Analysis of Energy Efficiency of Compressive Sensing in Wireless Sensor Networks”, IEEE Sensors Journal. Vol.13, No.5, 2015.

[7] J.S. Leu, T.H. Chiang, M.C.Yu and K.W. Su, “Energy Efficient Clustering Scheme for Prolonging the Lifetime of Wireless Sensor Network with Isolated Nodes”, IEEE Communications Letters, Vol.19, No.2, 2015, pp.259-262.

[8] X.Y. Liu, Y. Zhu and L. Kong, C. Liu, Y. Gu, Vasilakos, and M.Y. Wu, “CDC: Compressive Data Collection for Wireless Sensor Networks”, IEEE Transactions on Parallel and Distributed Systems, Vol.26, No.8, 2014, 2015, pp. 2188-2197.

[9] C. Luo, J. Sun and F. Wu, “Compressive Network Coding For Approximate Sensor Data Gathering”, in Proc. IEEE Global Telecommunications Conference (GLOBECOM), 2015, pp.1-6.

[10] Ming Xiao, Rong Du, “Lifetime Maximization for Sensor Networks with Wireless Energy Transfer”, IEEE International Conference on Communications, 2016, pp.20-25.

[11] Ming Xiao, Rong Du, “On Maximizing Sensor Network Lifetime by Energy Balancing”, IEEE Transactions on Computer and Communications, Accepted for Publications, 2017.

[12] Z.-H. Zhan, J. Zhang, Y. Li, and Y.-H. Shi, ''Orthogonal Learning Particle Swarm Optimization'', IEEE Trans. Evol. Comput., Vol. 15, No.6, Dec. 2011, pp. 832-847.

[13] Y. Valle, G. K. Venayagamoorthy, S. Mohagheghi, J. C. Hernandez, and R. G. Harley, ''Particle Swarm Optimization: Basic Concepts, Variants and Applications in Power Systems'', IEEE Trans. Evol. Comput., Vol.12, No.2, Apr. 2008, pp.171-195.

[14] Rongu Du, "Energy Efficient Monitoring of Water Distribution Networks via Compressive Sensing", IEEE Journal on Selected Areas in Communications, Vol.33, No.12, 2015.

[15] R. Du, L. Gkatzikis, C. Fischione and M. Xiao, "Energy Efficient Monitoring of Water Distribution Networks Via Compressive Sensing", in Proc. IEEE International Conference on Communications (ICC), 2015

[16] J.H. Chang and L. Tassiulas, "Maximum Lifetime Routing in Wireless Sensor Networks", IEEE/ ACM Trans. Networking (TON), Vol.12, No.4.

[17] A. Jarry, P. Leone, O. Powell and J. Rolim, "An Optimal Data Propagation Algorithm for Maximizing the Lifespan of Sensor Networks", in Distributed Computing in Sensor Systems. Springer, 2006, pp. 405-421.

[18] C. Cassandras, T. Wang and S. Pourazarm, "Optimal Routing And Energy Allocation For Lifetime Maximization Of Wireless Sensor Networks With Nonideal Batteries", IEEE Trans. Control of Network Systems, Vol.1, No.1, pp.86-98.

Automated Welding Torch Nozzle Cleaners

R. Nandha Kumar, R. Ohm Sakthivel, P.J.Guru kailash and A. Madhan Raj
Department of Mechatronics Engineering,
Agni College Of Technology, Chennai - 600 130, Tamil Nadu
E-mail: nandhakumar.mecha@act.edu.in, ohmsakthivel.mecha@act.edu.in

Abstract

In our modern world, the robots are used in industries to reduce human work and continuous production. For that industries has to spend more for buying robots for the production. Our aim is to make the production areas with minimizing human interference after the installation. The robotic arm is given some rest which increases the life time of the robot and to increase the production high (especially in MIG, TIG welding robots). This rest is given to the arm during the cleaning of torch nozzles. To achieve this process manual welding torch nozzle cleaning stations are modified in such a way that it is adaptable to any workplace.

Keywords: MIG-Metal Inert Gas welding, TIG-Tungsten Inert Gas welding.

1. INTRODUCTION

To achieve the stability, strength, durability, in the joining the two materials permanent joints are used widely in the industries. In olden days welding are done by humans which causes defect welding. In later 20th centuries due to the increased production requirement the industries need to produce more products.

After industrial revolution many automobile, OEM industries need more production. Thus the industries are automated with some robots which works without any tired. They must be programmed only once for the continuous and same repeating operations.

One of such operation is welding. The welding requires skilled labor work especially in automobile industries. Even though the skilled labor works some defects are there due to the mistake and carelessness of labors.

To reduce such kind of errors and mistakes the automatic robots are usedin industries. In recent days most of the automobile industries uses robotic works in the manufacturing especially in welding as shown in figure1.

Fig.1 Robot welding

2. MANUAL WELDING TORCH NOZZLE CLEANING STATION

The welding torch nozzle contains some slag sediment formation due to the fusion of filler material and the gas due to high temperature in the nozzle.

This sediments develops due to the continuous welding process and periodic resting strengthens the sediments in the nozzle and make it brittle shown in figure 2 and 3. This sediments makes the nozzle to be closed completely causes welding defects and the nozzle gets erode easily and quickly.

In order to reduce the risk of defects and the nozzle erosion automatic welding stations are used.

It requires continuous human interference to operate the manual machine each and every time of cleaning process in a industry.

Fig.2 Cleaning station

Fig.2 Slag formation in nozzle

Fig.3 Degradation of nozzles

3. AUTOMATED WELDING TORCH NOZZLE CLEANING STATION.

The risk of human intervention and is avoided and continuous production can be achieved by this type of automated stations. This stations reduce the axis movements of the robot and increases the life time to 1 year. This make the industries more profitable and cost efficient. The system is compact and kept near the welding robots shown in figure 4.

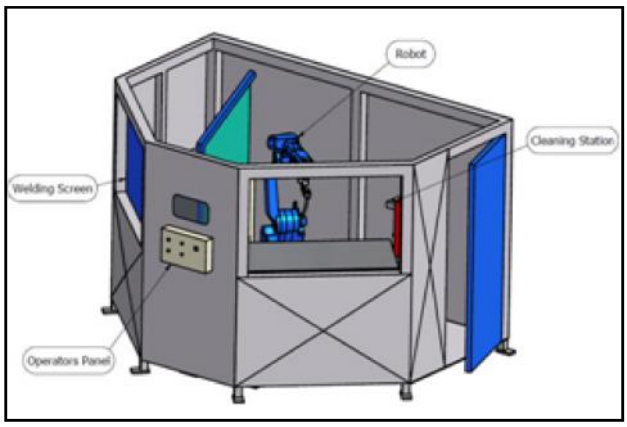


Fig.4 Welding area

4. SYSTEM DESCRIPTION

This setup consists of three major components in it to make the nozzle cleaning process effective.They are

- Reamers
- Anti-spatter sprays
- Wire cutters

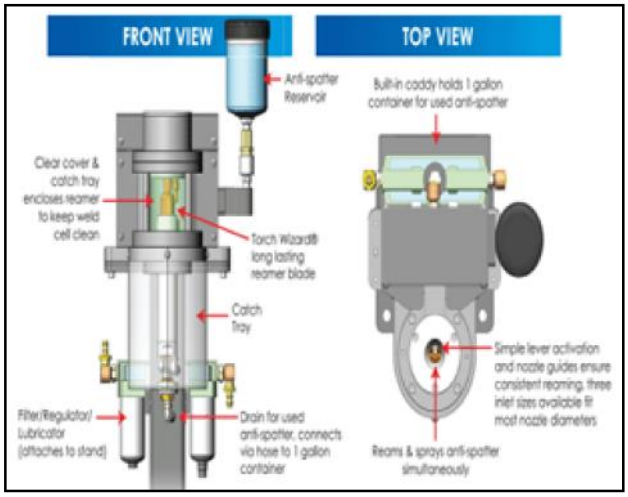


Fig.5 Automated cleaning station

4.1 Welding Reamers

It is used to clean the slag formation inside the nozzle. It is coupled with electric motor.By the relative motion produced between the reamer and nozzle the sediments are removed.

Fig.6 Reamers

4.2 Anti-Spatter

Heavy duty dry thin film welder anti spatter SPRAY USED for prevents spatter from adhering to weld beads, surrounding metal surfaces & weldingtips during welding.

Areas of Use:

Resistance welding tips,
MIG and TIG tips.

Fig.7 Anti-spatter liquids

4.3 Wire Cutter

It removes the ball at the end. The welding wire to provide smooth arc starts and consistent wire stick-out.

Fig.8 Wire cutters

5. MODIFICATIONS

- The above mentioned components are meshed in single trolley with timing controlled drives.
- The end effector home position is set to the station axis. The robot moves to the cleaning station for cleaning periodically after finishing jobs and send a feedback to the controller of the cleaning station.
- Un till all the cleaning process gets over the robot remains in idle state.
- The cleaning components are set up in a circular palette which acts in reciprocating motion and rotary motion.
- The anti-spatter is kept separate and it is connected with the help of duct.
- After the cleaning process is completed and completed feedback is given by the cleaning station controller to the welding robot so, that it can proceed its welding operation.

6. CONCLUSION

Thus by changing such modifications the following advantages can be achieved.The machine gets a rest time of 25secs max.Due to periodic resting no additional rest is required.It increases lifetime of motor and improve the working efficiency.Cleaning time program can be reduced.Human intervention can be reduced. Production can be increased.

REFERENCES

[1] Based on the Direct View and Study of Problems Faced by the Industries Which Are Have Using this Type of Automatic Welding Machines.

[2] Requirement of the Integrators and Industries like Hyundai, Fanuc, Ashok Leyland, BMW etc.

RFID Automated Retail Trolley with Ultrasonic Sensor

R. DeepanChakkaravarthi, G. Poovarasan, S. Mohamed Niyaz, Mahendran and T.R. Arunprasand and D.R. P. Rajarathnam
Department of Mechatronics, Paavai Engineering College(Autonomous), Namakkal-637 018, Tamil Nadu
E-mail: arunprasandramasamypec@paavai.edu.in

Abstract

Consider any shopping mall we have to wait much time for billing even though you purchase little things and we are not aware of cost of the product that we wanted to purchase. With the help of this project we reduce the billing time and customers can know the exact cost of the products that they purchased before billing so that they can do their shopping within their budget.The objective of this project is to improve the speed of purchase by using RFID. This project is designed to use the RFID based security system application in the shopping trolley. This project is used in shopping complex for purchase the products. In this project RFID Keyfob is used as security access for product. If the product is put in to the trolley means it will shows the amount and also the total amount. But in this project RFID Keyfob with 13.6 MHZ is used for accessing the products. So this project improves the security performance and also the speed. In this project RFID Keyfob is used as security access. So each product has the individual RFID fobs which represents the product name. RFID reader is interfaced with micro controller. Here the micro controller is the flash type re-programmable micro controller in which we already programmed with card number.RFID transceivers can only detect with minimum range so ultrasonic sensors ac be used within the area in order to improve the security performance at cheap price.

Keywords: *Frequency identification, Object detection ,Object counting, RFID,Key fobs*

1. INTRODUCTION

Waiting to pay at the checkouts at hypermarkets is very tiresome as people lead very busy lives. Therefore, waiting time should be managed and controlled. Checkout management is the next big technology for retailers in the modern world with less time spent queueing and better customer care. As customers hate long queues which leads to customer dissatisfaction in retailing . This arises whenever a shared facility needs to be accessed for service by a large number of customers. The expectation of short checkout queues is key ways to build customer loyalty and encourage spending . Historical Checkout Management systems were all about getting the customer in and out of your business as quick as possible in an orderly fashion. Today, Checkout Management systems can offer a lot more with the aid of advanced technology and development. This can help to improve business process and the overall customer experience. People are quite sensitive to wait times. There is no doubt that using a modern queue management system will bring improved efficiencies into the organization, through better understanding of customer expectations, greater opportunity to persuade customer to buy more products, better understanding of staff activity and better visibility of the business. The proposed solution for this problem is an intelligent queue management system which serves customer as well as the supermarket with minimum waiting time by managing the related resources such as staff, customers and checkouts efficiently and effectively.

2. LITERATURE REVIEW

Intelligent Retail Checkout Management System can be compared with other related products or systems which are available today, such asNextiva Queue Management, Qtech, QueueingSystem, EQMS, Qmatic, Irisysand AQMS-I6 .Irisysbasically uses non-instructive infrared sensors at store entrance and above the checkout lanes to monitor customer numbers and queueing behavior. The system is able to automate the capture of accurate data and calculate in real time the average queue length, average wait times, cashier idle times and overall transaction service time. Qmatic helps to organize the queues by providing visitors with virtual and linear queueing solutions, booked appointments or more sophisticated methods like mobile apps and SMS messages. AQMS-I6 is basically a token management

system. According to customer tokens and service counters the system manage waiting and reduce rush in counters. AQMS's goal is to reduce real and apparent waiting time, speeds up service delivery, improves service quality and increases customer satisfaction. In QtechQueueingSystem a keypad will be placed at the counter to facilitate the calling of the queue number by the counter staff which can be called in a sequential or random manner. Display panel shows both queue numbers and its corresponding counter number either with or without directional arrows. E-QMS is a method of queue management which is characterized by the use of electronic devices to manage the flow of customers or persons waiting in line to be served. A major feature of this type of queue management system is that, there is always an audio and/or flash light alert to let the next person know that he/she is ready to be served. Most of the above mentioned systems are recognized as token management systems where their goal is to manage waiting and reduce rush in counters.

3.METHODOLOGY

In the proposed system, after selecting the goods the customer will be requested to place his/her trolley in the trolley volume detection stall,the system calculates the number of goods in the trolley approximately.

Mean time system will keep track of the real time rush in each checkout. Then the system will decide the best checkout with minimum waiting time for each trolley and it will display to the relevant customer at the trolley volume detection stall. By analyzing the real time rush in checkout area, system calculates the average waiting time and displays it outside to attract more customers to the supermarket. In the entrance system tracks the number of customers entered and left the supermarket which calculates the current number of people inside the supermarket. In order to manage the staff efficiently, the system will provide predictions for future cashier formations and monthly reports which will be used to predict future sales.There by savingthe time and improving the shopping experience of the customers. The Block diagram is mentioned in fig (1) represents the postulated method.

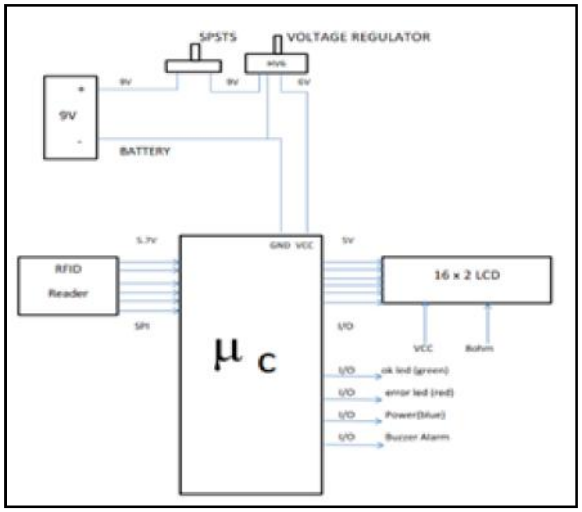


Fig.1 Trolley system block diagram

3.1 Components Required

3.1.1 a) ARDUINO 3.0

The Arduino Nano is a small,complete, and breadboard-friendly board based on the ATmega328 (Arduino Nano 3.0) or ATmega168 (Arduino Nano 2.x). It has more or less the same functionality of the Arduino Duemilanove, but in a different package. It lacks only a DC power jack, and works with a Mini-B USB cable instead of a standard one. The Nano was designed and is being produced by Gravitech.

The Arduino Nano is a small, complete, and breadboard-friendly board based on the ATmega328 (Arduino Nano 3.0) or ATmega168 (Arduino Nano 2.x). It has more or less the same functionality of the Arduino Duemilanove, but in a different package. It lacks only a DC power jack, and works with a Mini-B USB cable instead of a standard one. The Nano was designed and is being produced by Gravitech.

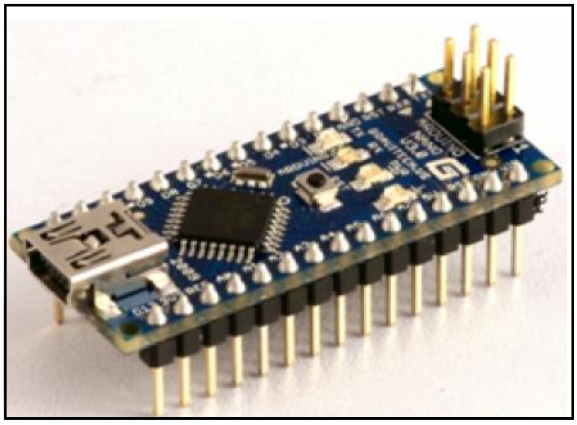


Fig.2 Arduino Nano

The Arduino Nano(Fig 2) can be powered via the Mini-B USB connection, 6-20V unregulated external power supply (pin 30), or 5V regulated external power supply (pin 27). The power source is automatically selected to the highest voltage source. The FTDI FT232RL chip on the Nano is only powered if the board is being powered over USB. As a result, when running on external (non-USB) power, the 3.3V output (which is supplied by the FTDI chip) is not available and the RX and TX LEDs will flicker if digital pins 0 or 1 are high.

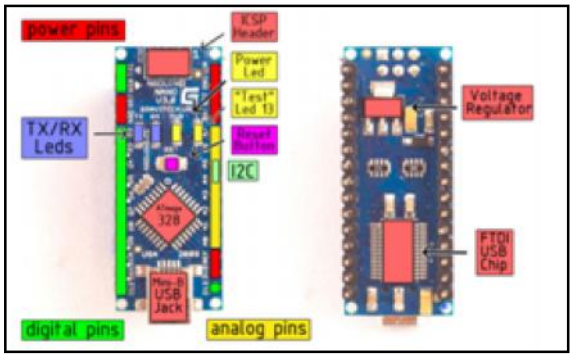


Fig.3 Arduinonano 3.0 with AT mega 328 processor

The ATmega168 has 16 KB of flash memory for storing code (of which 2 KB is used for the bootloader); the ATmega328 has 32 KB, (also with 2 KB used for the bootloader). The ATmega168 has 1 KB of SRAM and512 bytes of EEPROM (which can be read and written with the EEPROM library); the ATmega328 has 2 KB of SRAM and 1 KB of EEPROM

Specialized Functions:

- · Serial: 0 (RX) and 1 (TX).
  Used to receive (RX) and transmit (TX) TTL serial data. These pins are connected to the corresponding pins of the FTDI USB-to-TTL Serial chip.
- External Interrupts: 2 and 3.These pins can be configured to trigger an interrupt on a low value, a rising or falling edge, or change in value. See the attachInterrupt() function for details.
- PWM: 3, 5, 6, 9, 10, and 11.Provide 8-bit PWM output with the analogWrite() function.
- SPI: 10 (SS), 11 (MOSI), 12 (MISO), 13 (SCK).
  These pins support SPI communication, which, although provided by the underlying hardware, is not currently included in the Arduino language.
- LED: 13. There is a built-in LED connected to digital pin 13. When the pin is HIGH value, the LED is on, when the pin is LOW, it's off.

3.1.1 b) MFRC-522 RFID NFC Reader with Card and Tag

The module is based on RF module RC522 near field communication module. With operating frequency of 13.66Mhz where you can read and write a tag. Compatible in all gizDuino/ArduinoMicrocontroller boards.RFID tags support a larger set of unique IDs than bar codes and can incorporate additional data such as manufacturer, product type and even measure environmental factors such as temperature. Furthermore, RFID systems can discern many different tags located in the same general area without human assistance.

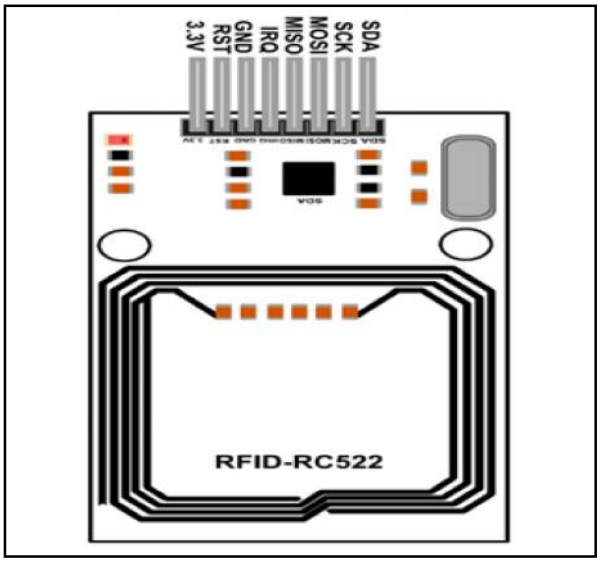


Fig.4 Sectional view of RFID transceiver

Fig.5 RFID receiver and tag

The general specifications of the RFID tag

Input Supply Voltage :3.3 VDC
Working Current :13 to 26mA
Part /number :MF522-ED
Card reading distance :0 to 60mm
Interface :SPI communication

Data Communication speed :10Mbit/s Max.
Operating Frequency :13.56Mhz
Supported card types : Mifare1 S50, Milfare1 S70, MifareUltraLighMifare Pro, MifareDesfire
Weight :8g
Dimensions :60mm x 40mm

3.1.1 c) Ultrasonic Sensor

Arduino Ultrasonic Range Detection Sensor with Arduino in order to calculate distances from objects. In this case I'm also altering the output of an LED with PWM according to how close an object is to the sensor. So the nearer you are the brighter the LED. So if we start with the Arduino Ultrasonic Range Detection Sensor, it's an IC that works by sending an ultrasound pulse at around 40Khz with a good ranging frequency.

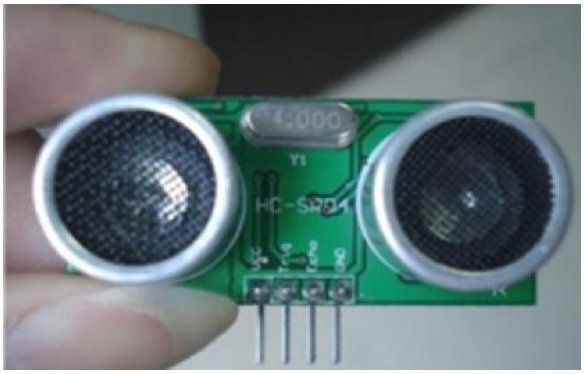


Fig.6 Ultrasonic sensor

It then waits and listens for the pulse to echo back, calculating the time taken in microseconds (1 microsecond = 1.0 Ч 10-6 seconds). You can trigger a pulse as fast as 20 times a second and it can determine objects up to 3 metres away and as near as 3cm. It needs a 5V power supply to run. Adding the Arduino Ultrasonic Range Detection Sensor to the Arduino is very easy, only 4 pins to worry about. Power, Ground, Trigger and Echo. Since it needs 5V and Arduino provides 5V I'm obviously going to use this to power it. Below is a diagram of my Arduino Ultrasonic Range Detection Sensor, showing the pins.

3.1.1 d) Voltage Regulator

The MC78XX/LM78XX/MC78XXA series of three terminal positive regulators are available in the TO-220/ D-PAK package and with several fixed output voltages, making them useful in a wide range of applications. Each type employs internal current limiting, thermal shut down and safe operating area protection, making it essentially indestructible. If adequate heat sinking is provided, they can deliver over 1A output current.

Although designed primarily as fixed voltage regulators, these devices can be used with external components to obtain adjustable voltages and currents.

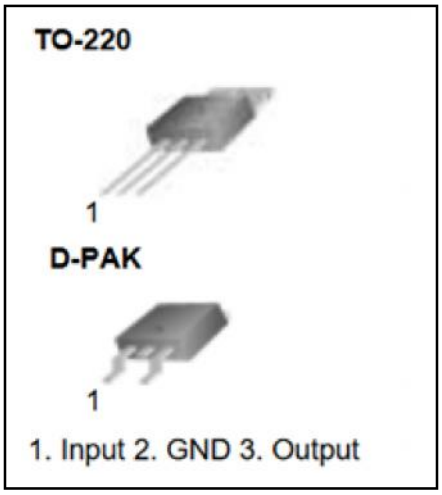


Fig.8 Voltage regulator

Features of the regulator are listed as follows,

- Output Current up to 1A
- Output Voltages of 5, 6, 8, 9, 10, 12, 15, 18, 24V
- Thermal Overload Protection
- Short Circuit Protection
- Output Transistor Safe Operating Area Protection

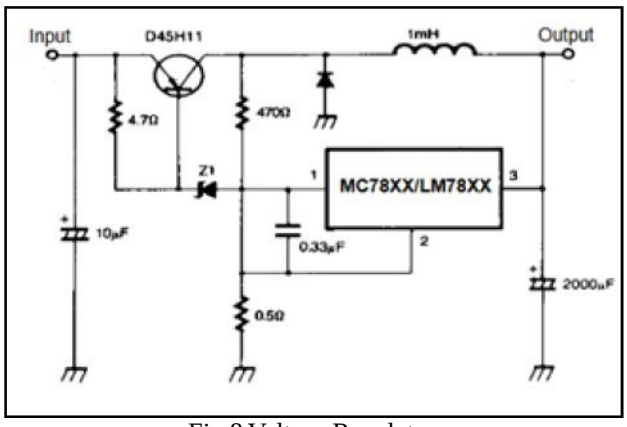


Fig.8 Voltage Regulator

3.1.1 e) Buzzer

This is a active buzzer so its easy in use. You just apply voltage to the buzzer and it makes sound. Disadvantage is that you can't determine the frequency of the sounds, for this you need a passive buzzer.
Schematic:

Module pin - = GND Module pin S = +5V
Connection to Arduino0 to digital pin 8 —> Module pin S
Arduino GND —> Module pin

Fig.9 Buzzer

3.1.1 f) LCD

LCD (Liquid Crystal Display screen is an electronic display module and find a wide range of applications. A 16x2 LCD display is very basic module and is very commonly used in various devices and circuits. These modules are preferred over seven segments and other multi segment LEDs.

The reasons being: LCDs are economical; easily programmable; have no limitation of displaying special & even custom characters (unlike in seven segments), an imitations and so on. A 16x2 LCD means it can display 16 characters per line and there are 2 such lines. In this LCD each character is displayed in 5x7 pixel matrix. This LCD has two registers, namely, Command and Data.

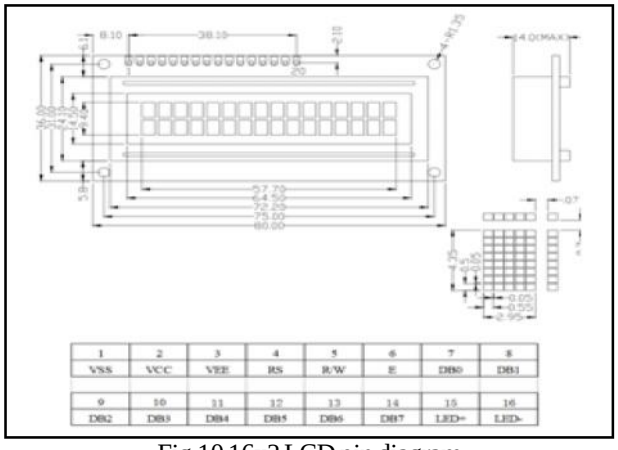


Fig.10 16x2 LCD pin diagram

Fig.11 16X2 LCD

The command register stores the command instructions given to the LCD. A command is an instruction given to LCD to do a predefined task like initializing it, clearing its screen, setting the cursor position, controlling display etc. The data register stores the data to be displayed on the LCD. The data is the ASCII value of the character to be displayed on the LCD.

3.2 Procedure

3.2.1 Object detection

The antenna emits radio signals to activate the tag and read and write data to it. Antennas are the conduits between the tag and the transceiver, which controls the system's data acquisition and communication. Antennas are available in a variety of shapes and sizes; they can be built into a frame to receive tag data from things passing through the transceiver required, a sensor device can activate thefield.When an RFID tag Passes through the electromagnetic zone, it detects the reader's activation signal.

3.2.2 RFID transceiver

The reader decodes the data encoded in the tag's integrated circuit (silicon chip) and the data is passed to the host computer for processing. The majority of RFID tags contain at least an integrated circuit for modulating and demodulating radio frequency and an antenna for transmitting and receiving signals. Frequency ranges vary from low frequencies of 125 to 134 kHz and 140 to 148.5 kHz, and high frequencies of 850 to 950 MHz and 2.4 to 2.5 GHz. Wavelengths in the 2.4 GHz range are limited because they can be absorbed by water.

3.2.3 Tag reader

RFID reader is used to read the data present in the RFID tag. RFID readers or receivers are composed of

a radio frequency module, a control unit and an antenna to interrogate electronic tags via radio frequency (RF) communication. Many also include an interface that communicates with an application. Readers can be hand held or mounted in strategic locations so as to ensure they are able to read the tags as the tags pass through an "interrogation zone." RFID systems can be classified by the type of tag and reader. A Passive Reader Active Tag (PRAT) system has a passive reader which only receives radio signals from active tags (battery operated, transmit only).

3.2.4 Power Supply

Power supply is used to give the 5V to the controller. 5V can be received from IC voltage regulator. Inside supply rectifier, filter is present.

3.2.5 Data Reading

This library makes it easy to use a Graphical LCD with Arduino this is an extensive modification of the ks0108 library that has higher performance, more features, supports more Arduino boards and is easier to integrate with different panels. Sketches written for the old library should work with little or no modification. Theconfigurationmechanism has been changed to facilitate use with a broad and AT megaControllers, See the section on sketch migration for details on modifications for the new library.

4. PURPOSE OF THE PROJECT

The objective of this project is to improve the speed of purchase by using RFID. This project is designed to use the RFID based security system application in the shopping trolley. This project is used in shopping complexshows the amount and also the total amount. But in this project RFID card is used for accessing the products. So this project improves the security performance and also the speed. The trolley developed will also have the provision to take out the printout the bill of the purchased materials which will be designed using .net graphical user interface with Access database. It will overcome the Barcode technology which gets lots of problems that will recover in this technology such as the barcode method is so slow and some time it will creating error at the reading the barcode if in case of damaged the barcode it won't be recognized the barcode tag by barcode reader. Since RFID has only a minimum frequency range an ultrasonic sensor can be used within the circumference area of the trolley. If not detected by the RFID the ultrasonic sensor senses the object within the area and buzzer a red light. Indicating that the customer or buyer is trying to steal the product.

4.1 Abbreviations and Acronyms Used

[1] RFID :- Radio Frequency Identification
[2] DoD :- Department of Defense
[3] EAS :- Electronic Article Surveillance
[4] EPC :- Electronic Product Code
[5] ISO :- Indian Standard of Organization
[6] ARPT :- Active Reader Passive Tag
[7] LCD :- Liquid Crystal Display
[8] PCB :- Printed Circuit Board
[9] BAP :- Battery Assisted Passive
[10] PRAT :- Passive Reader Active Tag
[11]IDE :- Integrated Development Environment

5.CONCLUSION

The intended objectives were successfully achieved in the prototype model developed. The developed product is easy to use, low-cost and does not need any special training. This project report reviews and exploits the existing developments and Different types of radio frequency identification technologies which are used for product identification, billing, etc. We have also learned the architecture of the system that can be used in the shopping systems for intelligent and easy shopping in the malls to save time, energy and money of the consumers. Present trends point towards the fast growth of RFID in the next decade. With around 600 million RFID tags sold in the year 2005 alone, value of market including systems, services and hardware is likely to grow by factor of 10 between years 2006 -2016. It is expected that total number of RFID tags delivered in the year 2016 will be around 450 times as compared to the ones delivered in the year 2006. This project reviews and exploits the existing developments and Different types of radio frequency identification technologies which are used for product identification, billing, etc. Thus the survey paper studies and evaluates research insight in Radio Frequency Identification systems from a big picture first. We have studied in detail about the business model, technological model and all related work and applications in the domain of RFID.

REFERENCES

[1] Nextiva_Queue_Management. [Online]. Available: http://www.verint.com/solutions/video Situation Intelligence/products/videousinessintelligence/next Queue Management Lindex. [Accessed :29-012016].

[2] QtechQueueingSystem. [Online].Available:http:// www.Queuein gsystem.com/submit-anarticle/our-featuredproducts.html. [Accessed: 10-0 1-2016].

[3] EQMS.[Online].Available:http:// www.academia.edu I7012630/ Design and Construction of an Electronic Queue Management system EQMS in Partial Fulfillment of the Award of Bachelor of Science B.Sc. Degree Chapter One Int [Accessed: IS-O 1-20 16]

[4] SatishKamble, SachinMeshram, Rahul Thokal, RoshanGakre. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, www.ijsce.org/attachments/File/v3i6/ F2040013614.pdf, Vol.3, No.6, January 2014.

[5] Ankit Anil Agarwal, Saurabh Kumar Sultania, GouravJaiswal, Prateek Jain. Control Theory and Informatics ISSN 2224-5774 (print) ISSN 2225-0492 (online) Vol 1, No.1, 2011 www.iiste.org/ Journals/index.php/CTI/article/download/697/5 90

[6] RFID Shopping System Senior Design 2011st. Cloud State University Aayush, Otabeck Atajanov, Hamad Alajamam www.slideshare.net/arttuladhar/ rfid-shopping-system.

[7] MikroElektronika http://www.mikroe.com/add-on-boards/display/easytft/ 5.LPC2148 Data sheet http://www.nxp.com/documents/data_sheet/ LPC2141_42_44 _46_48.pdf.

[8] F. Bielen, and N. Demoulin, "Waiting Time Influence on the Satisfaction-loyalty Relationship in Services", Managing Service Quality: An International Journal., Vol.17, No.2, 2007, pp.174-193.

[9] lrisys.{Online].Available http://www.irisys.net Queue Management. [Accessed:27-12-20IS]. Kotsis, G.(l992).

[10] MATlC.[Online].Available:http://www.qmatic.com I so I utions/bus iness-sol utionsl queue-anagementsystems/. [Accessed: 0301-2016].

[11] AQMS-16. [Online]. Available: http:// www.databyteindia.com/queue Management.html

[12] D. C. Tseng and C. H. Chang, "Color Segmentation Using Perceptual Attributes", InProc. of 11th lnternational Conference on Pattern Recognition, Amsterdam, HOLLAND, IAPR, IEEE, September 1992, pp.228-231.

[13] 1. H. Chen, S. Chang, "Learning Algorithms and Applications of Principal Component Analysis", Image Processing and Pattern Recognition, Chapter 1, C. T. Leondes, Academic Press, 1998.

[14] A. Tremeau and P. Colantoni, "Regions Adjacency Graph Applied to Color Image Segmentation", IEEE Transactions on Image Processing, 1998.

Implementation of Movie Recommendation System Using Multiple Users

V. Priyanka, R. Ragul, R. Ruqsana and V.Sivaranjani
Department of Computer Science, Kongu Engineering College, Erode - 638 060, Tamil Nadu

Abstract

Recommendation engine is a subclass of information filtering system that helps the users to predict ratings or to prefer an item under the consideration of the user. It is an intelligent system that helps a user to select an item among the interesting items. Movie recommendation systems are mainly focussed on collaborative filtering and clustering. Recommender systems is a method that enables filtering through large observation and information space in order to provide recommendations in the information space that user does not have any observation. This recommender system makes use of Alternating Least Square algorithm applied on Movielens dataset. This approach is compared with the existing approaches through evaluation parameters like standard deviation (SD) and root mean square error (RMSE) to deliver better results for movie recommender system. This approach offers minimum standard deviation and root mean square error. The final results obtained from the proposed approach on Movielens dataset provides high accuracy for reliability, efficiency and provides personalized movie recommendations comparing with existing methods.

Keywords: *Alternating least square, Collaborative filtering, K-means clustering, Machine learning, Recommendation system.*

1. INTRODUCTION

1.1 Machine Learning

Machine learning is a scientific discipline that explores the construction and study of algorithms that can learn from data. Such algorithms operate by building a model based on inputs and using that to make predictions or decisions, rather than following only explicitly programmed instructions. It has strong ties to artificial intelligence and optimization, which deliver methods, theory and application domains to the field. Machine learning is employed in a range of computing tasks where designing and programming explicit, rule-based algorithms is infeasible.

1.2 Data Mining

Data mining is a knowledge discovery process by analyzing large amounts of data stored in data warehouses widely used in large databases for finding frequent patterns. Thus data mining can also be named as knowledge mining. Data mining software is one of number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational database.

1.3 Collaborative Filtering

Collaborative filtering (CF) is a technique commonly used to build personalized recommendations on the Web. In collaborative filtering, algorithms are used to make automatic predictions about a user's interests by compiling preferences from several users. It is the way of filtering or calculating items through the sentiments of other people. It first gathers the movie ratings given by individuals and then recommends movies to the target user based on like-minded people with similar tastes and interests in the past. Difference business and product needs and a variety of algorithms that could be used for recommender systems yielded a rich set of methods that could be used for recommendations.

2. EXISTING SYSTEM

2.1 K-Means Clustering

K-Means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition an observation into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. In practice, in the management of the independent variables with more than two variables that use the same to replace two element vectors (x1,x2) with the same positions to n element vectors (x1,x2,.....xn). To

determine the group in advance, a random selection process starting from the center of each group, then measured the distance during each of the data center group of the conditions of the nearest or minimum distance in the group data.

The typical k-means clustering method,

- Clusters the data into K groups where k is predefined.
- Select K points at random as cluster centers.
- Assign objects to their closest cluster center, according to the Euclidean distance function.
- Calculate the centroid or mean of all objects in each cluster.
- Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds. The formula use for cluster center (centroid).

Clustering begin from K=2,3,4,.....,until suitability, this case use Euclidian distance is the basis of the group results. This is called the sum of squared errors (SSE).

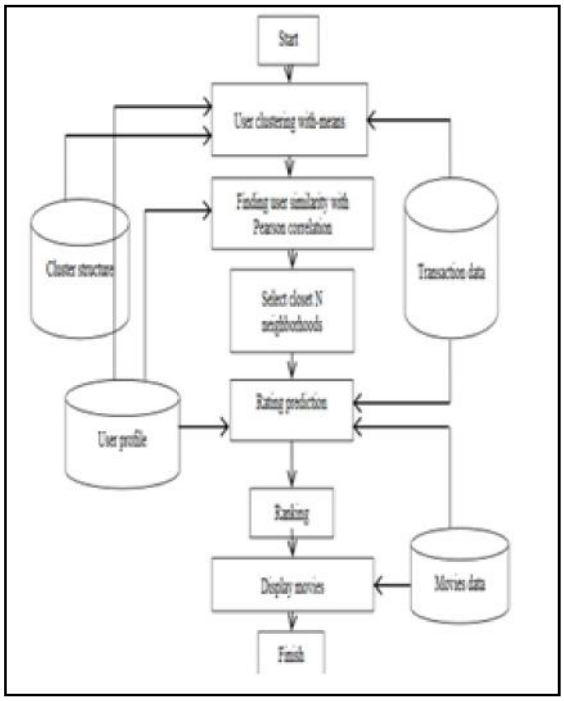


Fig.1 K-Means Clustering model for movie recommendation

Algorithm K-Means

Input: Dataset User-known Rating Matrix, number of clusters k

Output: Set of cluster centerss C,cluster membership vector S-Initialize cluster center C

-Do {Find closest cluster center for each user is User-item dataset using Euclidian distance and update S such that s1 is cluster ID of user1.Update C such that c1 is mean of users in ith cluster}

While (C doesn't change)

3. LITERATURE REVIEW

Recommendation systems have rigorously been used in various applications as a way to suggest items that a customer would likely be interested in by predicting customer preference. The most popular applications using recommendation systems are movies, music, news, grocery shopping, travel guides, online dating, books, restaurants, Ecommerce sites and so forth. Recommendation systems can be broadly categorized as contents-based filtering, collaborative filtering.

Guibing Guo, Jie Zhang, and Neil Yorke-Smith proposed a novel recommendation model regularized with user trust and item ratings it proposes TrustSVD, a trustbased matrix factorization technique for recommendations. Trust SVD assimilates several evidence sources into the recommendation prototype in order to diminish the data sparsity and cold start problems and their degradation of recommendation performance. A study of social trust data from four real-world data sets recommends that not only the obvious but also the contained influence of both ratings and trust should be taken into deliberation in a recommendation model.

The use of classification and regression tree (CART) was adopted by Amartya and Kundan [4] in their work. In constructing a decision tree, they applied both the gini index(g) and entropy value (ei) as the splitting indexes, the model was experimented with a given set of values, different sets of results were obtained for both the outlook, humidity, windy, Temp, and Time for execution. The result of the experiment shows that the best splitting attribute in each case was found to be outlook with the same order of splitting attributes for both indices.

Decision rule and Bayesian network, support vector machine and classification tree techniques were used by Rivas et al, to model accidents and incidents in two companies in order to identify the cause of accident. Data were collected through interview and designed. The experimental result was compared with statistics techniques, which shows that the Bayesian network and the other methods applied are more superior than the statistics technique.

4.PROPOSED SYSTEM

4.1Alternating Least Square

We have users u for items i matrix as in the following:
Qui={*r* is 0 if user u rate item 1, if user u did not rate item i}
where *r* is what rating values can be. If we have m users and n items, then we want to learn a matrix of factors which represent movies. That is, the factor vector for each movie and that would be how we represent the movie in the feature space. Note that, we do not have any knowledge of the category of the movie at this point. We also want to learn a factor vector for each user in a similar way how we represent the movie. Factor matrix for movies Y"Rfxn and factor matrix (each movie is a column vector) for users X"Rmxf(each user is a row vector). However, we have two unknown variables. Therefore, we will adopt an alternating least squares approach with regularization. By doing so, we first estimate Y using X and estimate X by using Y. After enough number of iterations, we are aiming to reach a convergence point where either the matrices X and Y are no longer changing or the change is quite small. However, there is a small problem in the data. We have neither user full data nor full items data, this is also why we are trying to build the recommendation engine in the first place. Therefore, we may want to penalize the movies that do not have ratings in the update rule. By doing so, we will depend on only the movies that have ratings from the users and do not make any assumption around the movies that are not rated in the recommendation. Let's call this weight matrix wui as such:
w_{ui}={0 if q_{ui}=0, 1 else
Then, cost functions for minimization:
$J(x_u)=(q_u-x_uY)W_u(q_u-x_uY)^T+\lambda x_u x^T u$
$J(y_i)=(q_i-Xy_i)Wi(q_i-Xy_i)^T+\lambda yiy^T i$

There is a need for regularization terms in order to avoid the overfitting the data. Ideally, regularization parameters need to be tuned using cross-validation in the dataset for algorithm to generalize better. Here using the whole dataset. Solutions for factor vectors are given as follows:
$x_u=(YW_uY^T+\lambda I)^{-1}YWuq_u$
$y_i=(X^TWiX+\lambda I)^{-1}X^TWiq_i$

where ,WudRnxn and WudRmxm diagonal matrices. The algorithm is pretty much of it. In the regularization, we may want to incorporate both factor matrices in the update rules as well if it would be more restrictive. The dataset is from MovieLens, contains 10000054 ratings and 95580 tags applied to 10681 movies by 71567 users.

4.2 Co-Clustering

Real world data is often bimodal, that is to say created by a joint interaction between two types of entities. For example, a user rating a document is affected by both the user characteristics and the document characteristics. Often, this type of signal is represented as a matrix, of which each dimension represents one of the entity types. Co-clustering is a term in data mining that relates to a simultaneous clustering of the rows and columns of a matrix. Co-clustering is extremely useful when the above mentioned pair wise interactions signal is sparse. The signal in the former example of users rating documents can be represented as a matrix with users as rows and documents as columns, and the inner cells of the matrix as the ratings, or any affinity signal of a user towards a document. Co-clustering this matrix can be explained as grouping both similar users and similar documents into, let's say, categories or interests, synchronously.

Alternating Least Squares rotates between fixing one of the unknowns ui or vj. When one is fixed the other can be computed by solving the least-squares problem. This approach is useful because it turns the previous non-convex problem into a quadratic that can be solved optimally. A general description of the algorithm for ALS algorithm for collaborative filtering is as follows:

Step 1: Initialize matrix V by assigning the average rating for that movie as the first row, and small random numbers for the remaining entries.
Step 2: Fix V, solve U by minimizing the RMSE function.
Step 3: Fix U, solve V by minimizing the RMSE function similarly.
Step 4: Repeat Steps 2 and 3 until convergence.

5. CONCLUSION AND FUTURE WORK

In conclusion the MovieLens dataset, which is extremely sparse, works much better on alternating least squares provides better scaling on this extremely sparse dataset. Based on the Movielens dataset, the experimental evaluation of the proposed approach proved that it is capable of providing high prediction accuracy and more reliable movie recommendations for users' preference comparing to the existing clustering-based CFs. The real objective of a recommender engine is to produce the best possible item recommendation lists for

each user, rather than just clustering users/item. The fact that MF is actually a soft clustering algorithm enables us to produce such recommendation lists. This recommender system makes use of Alternating Least Square algorithm applied on Movielens dataset. This approach is compared with the existing approaches through evaluation parameters like standard deviation (SD) and root mean square error (RMSE) to deliver better results for movie recommender system. This approach offers minimum standard deviation and root mean square error. The final results obtained from the proposed approach on Movielens dataset provides high accuracy for reliability, efficiency and provides personalized movie recommendations comparing with existing methods. As for cold-start issue, the experiment also demonstrated that our proposed approach is capable of generating effective estimation of movie ratings for new users via traditional movie recommendation systems.

As for future work, it can be continued to improve our approach to deal with higher dimensionality and sparsity issues in practical environment, and will explore more effective data reduction algorithms to couple with clustering-based CF. Furthermore, the study includes how the variation number of clusters may influence the movie recommendation scalability and reliability. To generate high personalized movie recommendations, other features of users, such as tags, context, and web of trust should be considered in our future studies.

REFERENCES

[1] Bartosz Kupisz Et. Al "Collaborative Filtering Recommendation Algorithm based on Hadoop and Spark", Published by IEEE 2015,pp 1510-1514 .

[2] B.M. Sarwar, G. Karypis, J. Konstan, J. Riedl, "Recommender Systems for Large-Scale E-Commerce: Scalable Neighborhood Formation Using Clustering", in Proceedings of International Conference on Computer and Information Technology, Dhaka, Bangladesh, 2002.

[3] Christina Christakou, Leonidas Lefakis, Spyros Vrettos and Andreas Stafylopatis; "A Movie Recommender System Based on Semi-supervised Clustering ", IEEE Computer Society Washington, DC, USA 2015.

[4] Costin-Gabriel Chiru, Vladimir-Nicolae Dinu , Ctlina Preda, Matei Macri ; "Movie Recommender System Using the User's Psychological Profile" in IEEE International Conference on ICCP, 2015.

[5] Dietmar Jannach, Gerhard Friedrich, "Tutorial: Recommender Systems", International Joint Conference on Artificial Intelligence, Beijing, August 4, 2013.

[6] Gaurangi, Eyrun, Nan; "MovieGEN: A Movie Recommendation System", UCSB.

[7] G. Adomavicius, A. Tuzhilin, "Toward the next Generation of Recommender System: A Survey of the State-Of-The-Art and Possible Extensions", IEEE Trans. Knowl. Data Eng., Vol.17, No.6, 2005, pp.734-749.

[8] G. Linden, B. Smith, J. York, Amazon.com recommendations: item to item collaborative filtering, IEEE Internet Comput., Vol.7, No.1, 2003, pp.76–80.

[9] Han J., Kamber M., "Data Mining: Concepts and Techniques", Morgan Kaufmann (Elsevier), 2006.

[10] Harpreet Kaur Virk, Er. Maninder Singh, "Analysis and Design of Hybrid Online Movie Recommender System", International Journal of Innovations in Engineering and Technology (IJIET), Vol.5, No.2, April 2015.

[11] IEEE paper on, "A Collaborative Filtering Recommendation Engine in A Distributed Environment", by Ghuli, P.; Ghosh, A.; Shettar, R.Contemporary.

[12] IEEE paper on Hybrid Recommendation System based on Collaborative Filtering and Fuzzy Numbers", by Miguel A. G. Pinto, Ricardo Tanscheit, Marley Vellasco Department of Electrical Engineering Pontical Catholic University of Rio de Janeiro Rio de Janeiro – Brazil 2012.

[13] N K. Jha, M.Kumar, A.Kumar and V.K.Gupta, "Customer Classification in Retail Marketing by Data Mining", International Journal of Scientific & Engineering Research, ISSN 22295518, Vol.5, No.4, April 2014.

[14] Ricci and F. Del Missier, "Supporting Travel Decision making Through Personalized Recommendation", Design Personalized User Experience for e-commerce, 2004, pp. 221-251.

[15] M. Steinbach, P.Tan, V. Kumar, "Introduction to Data Mining", Pearson, 2007.

Indian Journal of Engineering, Science, and Technology (IJEST)

(ISSN: 0973-6255)
(A half-yearly refereed research journal)

Information for Authors

1. All papers should be addressed to The Editor-in-Chief, Indian Journal of Engineering, Science, and Technology (IJEST), Bannari Amman Institute of Technology, Sathyamangalam - 638 401, Erode District, Tamil Nadu, India.
2. Two copies of manuscript along with soft copy are to be sent.
3. A CD-ROM containing the text, figures and tables should separately be sent along with the hard copies.
4. Submission of a manuscript implies that : (i) The work described has not been published before; (ii) It is not under consideration for publication elsewhere.
5. Manuscript will be reviewed by experts in the corresponding research area, and their recommendations will be communicated to the authors.

Guidelines for submission

Manuscript Formats

The manuscript should be about 8 pages in length, typed in double space with Times New Roman font, size 12, Double column on A4 size paper with one inch margin on all sides and should include 75-200 words abstract, 5-10 relevant key words, and a short (50-100 words) biography statement. The pages should be consecutively numbered, starting with the title page and through the text, references, tables, figure and legends. The title should be brief, specific and amenable to indexing. The article should include an abstract, introduction, body of paper containing headings, sub-headings, illustrations and conclusions.

References

A numbered list of references must be provided at the end of the paper. The list should be arranged in the order of citation in text, not in alphabetical order. List only one reference per reference number. Each reference number should be enclosed by square brackets.

In text, citations of references may be given simply as "[1]". Similarly, it is not necessary to mention the authors of a reference unless the mention is relevant to the text.

Example

[1] M.Demic, "Optimization of Characteristics of the Elasto-Damping Elements of Cars from the Aspect of Comfort and Handling", International Journal of Vehicle Design, Vol.13, No.1, 1992, pp. 29-46.
[2] S.A.Austin, "The Vibration Damping Effect of an Electro-Rheological Fluid", ASME Journal of Vibration and Acoustics, Vol.115, No.1, 1993, pp. 136-140.

SUBSCRIPTION

The annual subscription for IJEST is Rs.600/- which includes postal charges. To subscribe for IJEST a Demand Draft may be sent in favour of IJEST, payable at Sathyamangalam and addressed to IJEST. Subscription order form can be downloaded from the following link http:// www.bitsathy.ac.in./ ijest.html.

For subscription / further details please contact:

IJEST
Bannari Amman Institute of Technology
Sathyamangalam - 638 401, Erode District, Tamil Nadu Ph: 04295 - 226340 - 44
Fax: 04295 - 226666 E-mail: ijest@bitsathy.ac.in Web:www.bitsathy.ac.in

Indian Journal of Engineering, Science, and Technology

Volume 12, Number 1 January - June 2018

CONTENTS