JIT

Volume 5

Issue 1

2017-18

Journal of

Information Technology

Volume 5, Issue 1, 2017-18

PCE
PILLAI COLLEGE OF ENGINEERING

Department of Information Technology

Pillai College of Engineering

Plot No. 10, Sector 16, New Panvel - 410206
Maharashtra, India.

ACCREDITED WITH GRADE A+ NAAC

Journal of Information Technology (JIT)

JIT, Volume 5, Issue 1, 2017-18

Message

Dr. Sandeep M. Joshi
Principal, PCE

It is a matter of great pleasure that the Department of Information Technology, PCE New Panvel is bringing out regularly its issue of Department Journal - a creative hard research work of the students and faculty.

Thanks to the team for untiring efforts to inculcate strong values in our students. Values would help them to distinguish right from wrong and make the world a better place to live.

My best wishes for another wonderful year ahead.

Editorial

Dr. Satishkumar L Varma
Editor-in -Chief

Dear faculty and students of Pillai College of Engineerig,
Greetings!

It is with deep satisfaction that I applaud and congratulate you for contributing technical papers. I feel proud to bring out this issue of the Journal of Information Technology (JIT).

This journal focuses on a variety of topics such as Data Mining, Image Processing, Neural Networks, Machine Learning, Data Security and IoT. It explores the method of enhancing data security by using Hybrid Cryptography. Various Machine Learning applications are also covered such as product suggestions, image categorization and image description generation.

This issue covers twelve papers published by faculty and under-graduate students of Department of Information Technology, Pillai College of Engineering (PCE). I am happy to note that this issue of PCE JIT will be helpful for the future engineers working in the areas of Data Security, Machine Learning and Data Mining.

I wish a successful and fruitful publication life with our department journals.

We are honored to dedicate the issue of JIT to all the students and faculty of PCE.

Contents

Contents

About the Editors

Satishkumar L. Varma received his Ph.D degree in Computer Science and Engineering under the guidance of Dr. S N Talbar from SGGS I E & T, SRTMU, Nanded, India in March 2013. He received his graduation and postgraduation degree in Computer Engineering from Dr. BATU, Lonere, Raigad, MH, India, in the year 2000 and 2004, respectively. He is currently working as Professor and Head in the Department of Information Technology, Pillai College of Engineering, New Panvel, MH, India. He has twenty-one years of experience in teaching and research. He has received and successfully executed three R&D Funded Projects of amount more than Rs 9 Lakhs. He has published 1 copyrights, 8 Book Chapters, more than 29 refereed Journal papers and more than 32 papers in referred National as well as International Conferences including IEEE, Springer and IET with a second best paper award at National level paper presentation competition in Threshold–2000. He is recognized as a Teacher of University of Mumbai in Ph.D Degree in Computer Engineering. His delivered talks include Image Processing, Object Oriented Analysis and Design, MATLAB, Scilab, Hadoop, LaTeX, Android, Python, R, Google Scripts and Docs. He is a member of Technical Professional society in IEEE, ISTE, and CSI. His research interests involve Digital Image and Video Processing, Medical Imaging, AI and Machine Learning, Soft Computing, Data Mining and Information Retrieval.

Sushopti Gawade is pursuing Ph.D in Computer Engineering with research area Usability Engineering in Agriculture Domain. She has received B.E in Computer Science and Engineering 1997 and M E Computer Science and Engineering from Walchand College of Engineering Sangli in 2006. Currently she is working as a Professor in Pillai College of Engineering, Panvel. She is a highly dedicated and performance-driven professional with 21 years of teaching experience in Mumbai University. She has the ability to coordinate and direct all phases of project‑based efforts while managing, motivating, and leading the project team. She is an excellent problem solver and opportunities identifier to improve and resolve critical issues. She is a quick learner of new concepts and technologies and has excellent ability in expressing ideas clearly and good team management skills.

Gayatri Hegde has received her M.E in Computer Engineering from Pillai College of Engineering, Mumbai University. She has received M.B.A degree in Systems and Marketing from Sikkim Manipal University and completed B.E in Computer Science and Engineering from Basaveshwar Engineering College, University, Karnataka. She is currently working as assistant professor in Pillai College of Engineering, New Panvel, Maharashtra since 2010. She has 5 conference and journal publications and has attended 3 FDP. Her area of interest includes Operating system, Cloud Computing, Big Data Analytics and Distributed Systems.

Sagar Kulkarni received his M.E degree in Computer Engineering in 2014 from University of Mumbai, India. During his Masters and he received B.E in CSE in 2008 from Shivaji University, Kolhapur. He has participated in Avishkar project competition and he won a gold medal both at university level and state level. The North Maharashtra University has awarded him with a fellowship for his research work in Masters degree. He has more than 12 years of experience in teaching. He is currently working as Assistant Professor in the Computer Engineering department at Pillai College of Engineering, New Panvel, University of Mumbai, India. He has published more than 20 research papers in various national / International journals and conferences. He has actively contributed/ Conducted workshops/Training programs while in his tenure. His areas of interest are Natural Language Processing, Information Retrieval, System Programming and Compiler construction, Cyber Security, Digital Forensics etc.

ANALYSIS OF AMAZON DATA TO BOOST RETAIL REVENUE

Pillai College Of Engineering
Information Technology Department,Mumbai university

Shaikh Gulam Ahmed Raza
Information Technology Engineering
Mumbai, India
shaikh786028@gmail.com

Nishant Thakur
Information Technology Engineering
Mumbai, Thane
nishantthakur8590@gmail.com

Panhale Siddhesh
Information Technology Engineering
Mumbai, India
spanhale845@student.mes.ac.in

Prof. Madhu N.
Pillai College of Engineering,
Navi Mumbai,India
madhmn@mes.ac.in

***Abstract*—Recommendation system provides to understand a person's taste and desirable content for them automatically based on pattern between their likes and rating or different reviews. In this project, we have proposed a recommendation system for the large amount of amazon data in the form of rating, reviews, complaints and feedback about any product available on Amazon website using Hadoop framework. Apache Pig is used for analyzing Amazon data. Apache Pig is a platform for processing and analyzing large datasets. Data manipulation is performed with help of Apache Pig and Hadoop. Collaborative filtering is the prior choice of most recommendation services. The main idea behind collaborative filtering is that users having similar taste or opinion for particular item will also have a match for other items or services. Apache Hive is used as data Warehousing software used to analyze data with Hadoop on the basis of query summarization. Technology behind recommendation system is Hybrid Collaborative Filtering Agglomerative Clustering algorithm.**

I. INTRODUCTION

Data generated by E-Commerce sites and Consumer reviews are invaluable as a source of data to help people form opinions on a wide range of products. Beyond telling us whether a product is 'good' or 'bad', reviews tell us about a wide range of personal experiences; these include objective descriptions of the products' properties, subjective qualitative assessments, as well as unique use-(or failure-) cases. The value and diversity of these opinions raises two questions of interest to us:

(1) How can we help users navigate massive volumes of consumer opinions in order to find those
that are relevant to their decision? And

(2) how can we address specific queries that a user wishes to answer in order to evaluate a

Amazon offer community-Q/A systems that allow users to pose product specific questions to other consumers.1 Our goal here is to respond to such queries automatically and on-demand. To achieve this we make the basic insight that our two goals above naturally complement each other: given a large volume of community-Q/A data (i.e., questions and answers), and a large volumeof reviews, we can automatically learn what makes a review relevant to a query. We see several reasons why reviews might be a useful source of information to address product -related queries, especially compared to existing work that aims to solve Q/A-like tasks by building knowledge bases of facts about the entities in question:

_ General question-answering is a challenging open problem. It is certainly hard to imagine that
a query such as "Will this baby seat fit in the overhead compartment of a 747?" could be answered by building a knowledge-base using current techniques. However it is more plausible that some review of that product will contain information that is relevant to this query. By casting the problem as one of surfacing relevant opinions (rather than necessarily generating a conclusive answer), we can circumvent this difficulty, allowing us to handle complex and arbitrary queries. Fundamentally, many of the questions users ask on review websites will be those that can't be answered using knowledge bases derived from product specifications, but rather their questions will be concerned with subjective personal experiences. Reviews are a natural and rich source ofdata to address such queries. Finally, the massive volume and range of opinions makes review systems difficult to navigate,especially if a user is interested in some niche aspect of a product. Thus a system

that identifies opinions relevant to a specific query is of fundamental value in helping users to navigate such large corpora of reviews. Product review websites provide an incredible lens into the wide variety of opinions and experiences of different people, and play a critical role in helping users discover products that match their personal needs and preferences. To help address questions that can't easily be answered by reading others' reviews, some review websites also allow users to pose questions to he community via a question-answering (QA) system. As one would expect, just as opinions diverge among different reviewers, answers to such questions may also be subjective, opinionated, and divergent. This means that answering such questions automatically is quite different from traditional QA tasks, where it is assumed that a single 'correct' answer is available. While recent work introduced the idea of question -answering using product reviews, it did not account for two aspects that we consider in this paper:

(1) Questions have multiple, often divergent, answers, and this full spectrum of answers should somehow be used to train the system;

(2) What makes a 'good' answer depends on the asker and the answerer, and these factors should be incorporated in order for the system to be more personalized.

Here we build a new QA dataset with 800 thousand questions—and over 3.1 million answers—and show that explicitly accounting for personalization and ambiguity leads both to quantitatively better answers, but also a more nuanced view of the range of supporting, but subjective, opinions. Consumers have many choices, and e-commerce systems must provide appealing suggestions before impatient customers defect to competing websites. Due to increasing data volumes, diverse information sources and complex processing requirements, the retailer struggled to deliver speedy online computing..

(1) Apache Hadoop

Hadoop is a software framework that supports data-intensive distributed applications. It enables applications to work with thousands of computational independent computers and peta bytes of data. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Hadoop is completely written in Java and is cross platform. Hadoop enables the development of reliable, scalable, efficient, economical and distributed computing using very simple Java interfaces - massive parallel code without the pain.

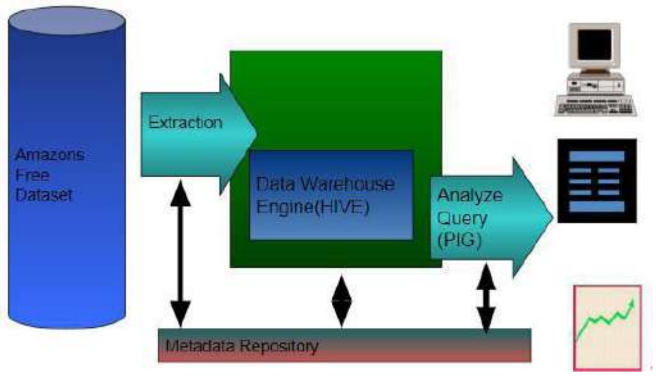


Fig: Apache Hadoop Farmewrok

(2) Hadoop Distributed File System (HDFS)

Hadoop includes a fault tolerant storage system called the Hadoop Distributed File System, or HDFS. HDFS is to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop is ideal for storing large amounts of data, like terabytes and petabytes, and uses HDFS as its storage system.

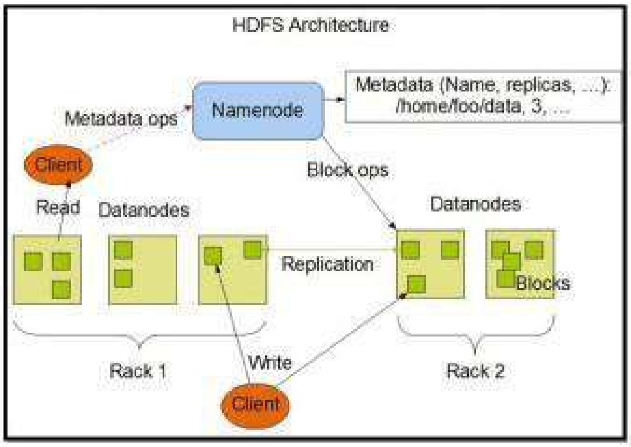


Fig: HDFS Architecture.

(3) MapReduce - Programming Model

MapReduce is a linearly scalable programming model.when

the size of the input data is doubled, a job will run twice as slow. But the size of the cluster is increased, a job will run as fast as the original one. Hadoop's MapReduce and HDFS use simple robust framework runs on commodity hardware to deliver high data availability and to analyze 20enormous amounts of information quickly. Hadoop offers enterprises a powerful new tool for handling big data. Hadoop creates clusters of machines and coordinates work among them. Clusters can be built with inexpensive computers.

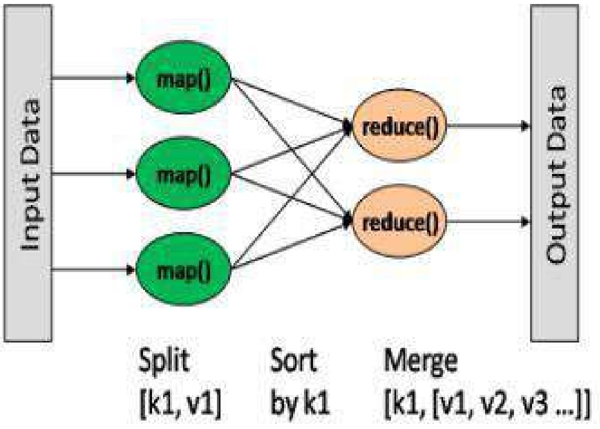


Fig: MapReduce Architecture.

(4) Apache Hive

Apache Hive is an open-source data warehouse system for querying and analyzing large datasetsvstored in Hadoop files. Hadoop is a framework for handling large datasets in a distributed 21computing environment. HiveQL is the Hive query language. Like all SQL dialects in widespread use, it doesn't fully conform to any particular revision of the ANSI SQL standard.

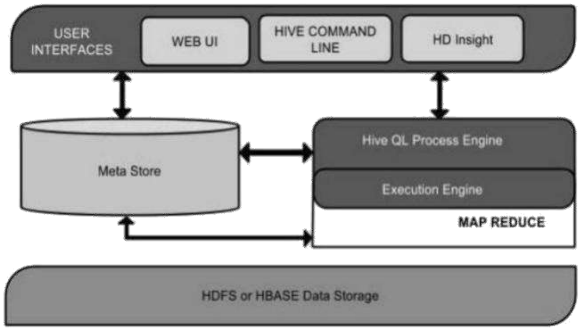
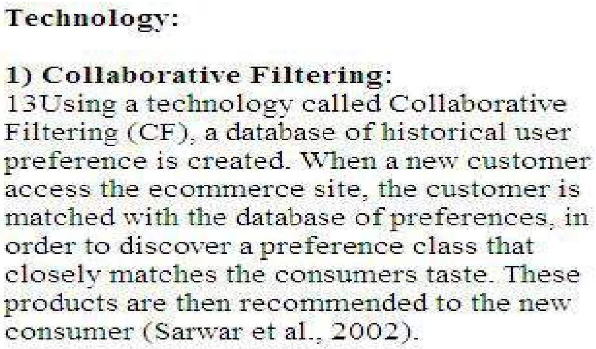


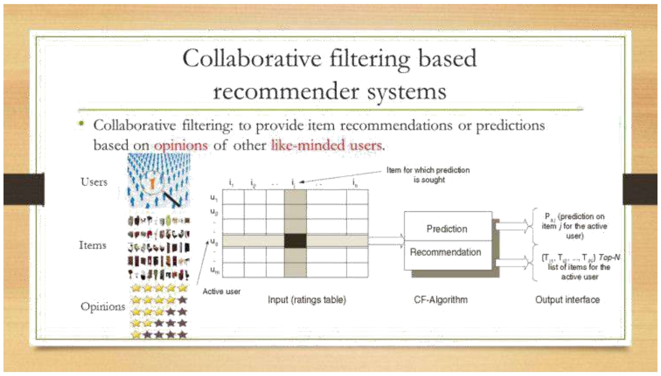
Fig: Apache Hive.



Technology:

1) Collaborative Filtering:

13Using a technology called Collaborative Filtering (CF), a database of historical user preference is created. When a new customer access the ecommerce site, the customer is matched with the database of preferences, in order to discover a preference class that closely matches the consumers taste. These products are then recommended to the new consumer (Sarwar et al., 2002).



Collaborative filtering based recommender systems

- Collaborative filtering: to provide item recommendations or predictions based on opinions of other like-minded users.

Users
Items
Opinions
Item for which prediction is sought
i_1 i_2 .. i_j .. i_n
u_1
u_2
u_a
u_m
Active user
Input (ratings table)
Prediction
Recommendation
CF-Algorithm
$P_{a,j}$ (prediction on item j for the active user)
$\{T_{i1}, T_{i2}, ..., T_{iN}\}$ Top-N list of items for the active user
Output interface

Figure : Collaborative Filtering Algorithm based on (Sarwar et al., 2002)

3.2 Clustering Algorithm

Clustering Algorithm technique works by identifying groups of users that have similar preferences. These users are then clustered into a single group and are given a unique identifier. New customers cluster are predicted by calculating the average similarities of the individual members in that cluster. Hence a user could be a partial member of more than one cluster depending of the weight of the user's average opinion (Sarwar et al., 2002)

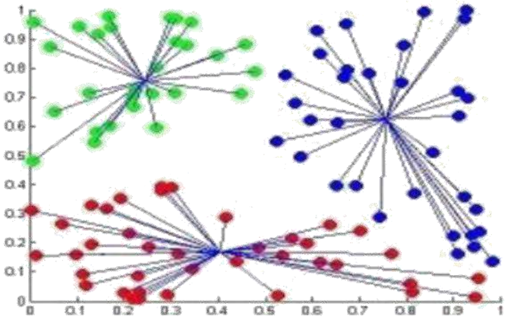
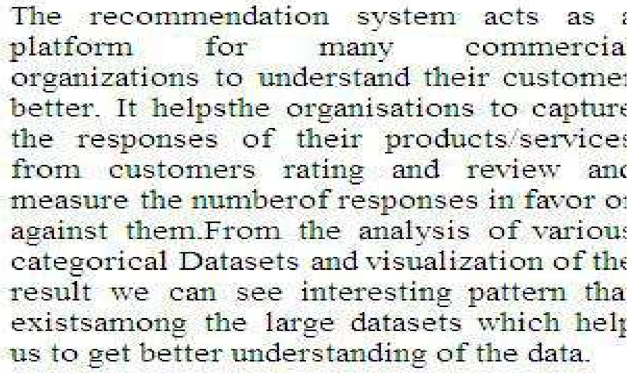


Figure: Clustering Algorithm based on (Sarwar et al., 2002)

CONCLUSION:



The recommendation system acts as a platform for many commercia organizations to understand their customer better. It helpsthe organisations to capture the responses of their products/services from customers rating and review and measure the numberof responses in favor or against them.From the analysis of various categorical Datasets and visualization of the result we can see interesting pattern tha existsamong the large datasets which help us to get better understanding of the data.

References

[1]Apache sqoop from official website: http://sqoop.apache.org/docs/

[2] Why Big Data is a must in E-Commerce", Guest post by Jerry Jao, CEO of Retention

available at http://www.bigdatalandscape.com/news/why-big-data-is-a-must-in-ecommerce .

[3] Gayathri Ravichandran, "Big Data Processing with Hadoop" Volume 4.81 ,Pages 448-451,

IRJET , Feb 17

[4] Prarthana Rao H M, " A Review on Big Data: Recommendation System" Volume 6 Issue 3

,Page 445-449 ,IJIET , Feb 2016.

[5] Khushboo R. Shrote and A.V. Deorankar, " Review based service recommendation for big data", Volume 4 ,Pages 742-756, AEEICB , Feb

2016.

[6] Vishal Nehe and Abhishek konduri , " Commercial Product Analysis Using Hadoop MapReduce" ,Volume 4.45, Pages 2429-2433, IRJET, April 2016

[7] Essa, Y.M., Attiya, G. & El-sayed, A., 2013.
Mobile Agent based A New Framework for Improving
Big DataAnalysis.
[8]Defination of sqoop. Available FTP:
https://www.javatpoint.com/what-is-sqoop
[10]Hive Introduction. Available
FTP:
https://www.tutorialspoint.com/hive/hive_introduction.htm
[11]Overview of the Collaborative Filtering Process.
Available
FTP:http://www10.org/cdrom/papers/519/node6.html
[12]Data Clustering Algorithm. Available

Automatic Image Description Generator Using Neural Networks

Ajit Tiwari, Student,PCE, Amar Roundhal, Student,PCE, and Krishnendu Nair, Faculty, PCE
Department of Information Technology
Panvel (Maharashtra), India

Abstract- Being able to automatically describe the content of an image using properly formed English sentences is a simple task for humans but our smartest systems are still not able to this simple task , but it could have great impact, for instance by helping visually impaired people better understand the content of images on the web.This task is significantly harder, for example, than the well-studied image classification or object recognition tasks, which have been a main focus in the computer vision community.We would like to present in this work a single joint model that takes an image as input, and is trained to maximize the likelihood of producing a target sequence of words where each word comes from a given dictionary, that describes the image adequately.

I. INTRODUCTION

Our Model uses an approach of using two neural networks for performing the task of image recognition and sentence formation.The both networks are Convolutional Neural Network[1][2] and Recurrent Neural Network[5].They both are connected in type of feedforward mechanism so the output from one layer is provided as input to other layers.

II. DIFFERENT TECHNIQUES OF IMAGE DESCRIPTION MODEL

a) Convolutional Neural Network
b) Recurrent Neural Network

A. Convolutional Neural Network:

A convolutional neural network (CNN, or ConvNet)[1] is a class of deep, feed-forward artificial neural networks that has successfully been applied to analyzing visual imagery.CNN's use a variation of multilayer perceptrons designed to require minimal preprocessing.CNN's use relatively little pre-processing compared to other image classification algorithms. This means that the network learns the features that in traditional algorithms were hand-engineered.This independence from prior knowledge and human effort in feature design is a major advantage.In the context of machine vision, image recognition is the capability of a software to identify people, places, objects, actions and writing in images. To achieve image recognition, the computers can utilise machine vision technologies in combination with artificial intelligence software and a camera.

While it is very easy for human and animal brains to recognize objects, the computers have difficulty with the same task. When we look at something like a tree or a car or our friend, we usually don't have to study it consciously before we can tell what it is. However, for a computer, identifying anything(be it a clock, or a chair, human beings or animals) represents a very difficult problem and the stakes for finding a solution to that problem are very high.Image recognition is a machine learning method and it is designed to resemble the way a human brain functions. With this method, the computers are taught to recognize the visual elements within an image.VGGNet[1] is a neural network that performed very well in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[3] in 2014.VGGNet has a 21 layer architecture which extracts the features of images.VGG16 is trained for classification of images into one of the 1000 object classes which is done after the image vector is passed through last softmax layer that converts the image vector into different probabilities indicating which class of object they belong to.For our model we are not interested in classifying the image but we need the internal representation of image in vector form just before applying softmax function.This vector is formed at last fully connected layer that gives us 4096 element vector.After we have acquired every feature vector we can use it as input to RNN network.

B. Recurrent Neural Networks

A recurrent neural network (RNN)[2][5] is a class of artificial neural network where connections between units form a directed cycle. This allows it to exhibit dynamic temporal behavior. Unlike feedforward neural networks, RNNs can use their internal memory to process arbitrary sequences of inputs.Long short-term memory (LSTM) are a special kind of RNN, capable of learning long-term dependencies suitable for our problem.All recurrent neural networks have the form of a chain of repeating modules of neural network.LSTMs also have this chain like structure, but the repeating module has a different structure.This makes our words to go through LSTM nodes again and again which makes the system understand positioning of the word as well as previous words.

In the given diagram, {s0, s1, ..., sN} represent the words of the caption we are trying to predict and {wes0, wes1, ..., wesN-1} are the word embedding vectors for each word. The outputs {p1, p2, ..., pN} of the LSTM are probability distributions generated by the model for the next word in the sentence. The model is trained to minimize the negative sum of the log probabilities of each word.

III. ARCHITECTURE

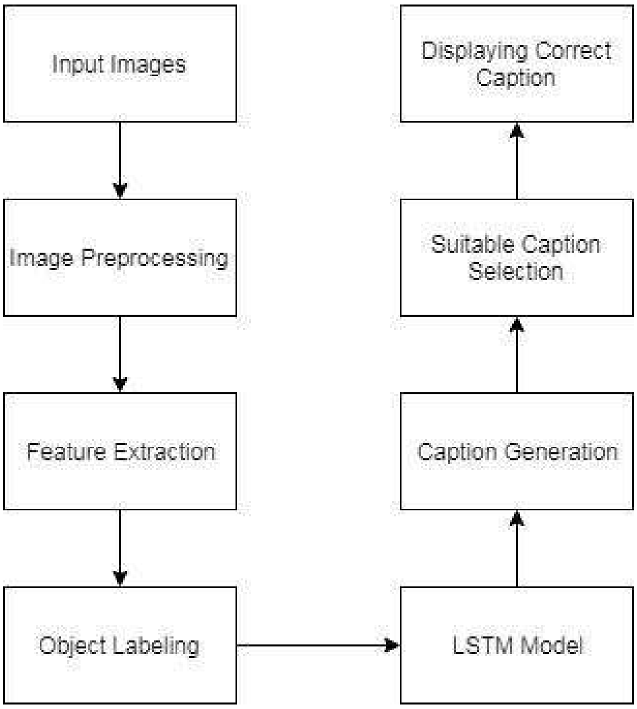


Figure 1: Architectural overview of Image Description Model

IV. IMPLEMENTATION

This outlines the specific steps taken by us in order to generate Image Description model

1)DataBase Creation:-

The first step in our process is to create our own database with few images and two description per images and we have also described set of training and development images that means of all the photos in our dataset we have chosen some images for training and other remaining images to verify the accuracy of the description generated while training,this helps us to choose the best model in various iterations.We have selected around 170 images and two different description per images we have vocabulary list of 89 words so that means when a new image is given to the model for generating description it has a choice of words of 89 words.

2)Image Feature Extraction:-

Next Step is to extract features of every image present in the database,these features can be extracted by passing the image through VGG16 convolution model.We have removed the last layer present in the Model which classifies the images into object classes.Once an image is passed through the model we get a NumPy array of 1x4096 dimensions.After every image is passed through the model we have saved all the arrays into a pickle file because we require these features later for training and we don’t want to extract features everytime we use images.

3)Word Encoding:-

After we have extracted features of every image we can now move on to working with words present in our database.In order to train the model it is necessary that we have to split the sentences into different words.The description text will need to be encoded to numbers before it can be presented to the model as in input or compared to the model’s predictions.

The first step in encoding the data is to create a consistent mapping from words to unique integer values.This can be done by using a hashing function that can convert every word into unique integer.This can be called as encoded text.Next we have to create sequences of input-output pairs for training and validation. The model will be provided one word and the photo and generate the next word. Then the first two words of the description will be provided to the model as input with the image to generate the next word. This is how the model will be trained.We also have to provide a starting token and ending token to our sentences in order for the model to understand where the descriptions are starting and ending these are given as ‘startseq’ and ‘endseq’.

4)Model Training:-

After we know how to create sequences of input-output pairings required for training and validation.We now have to design our neural network for training the model.This model will take image feature vector and encoded text as input and will learn which photo features match to which words and it will save that learned to a .h5 model.This model can be used for predicting descriptions of new images.The process of learning encoded text from photo features is done for various iterations called epochs.After every epoch we check whether the predicted text is accurate by checking with validation set and calculate the loss.After every epoch we check whether the loss is reduced and if it is then we save that model.We have to design the network is such a way that every neuron passes information to next neuron.Such type of network is called Feed-Forward network.

i)Photo Feature Extractor. This is a 16-layer VGG[1] model pre-trained on the ImageNet dataset. We have pre-processed the photos with the VGG model (without the output layer) and will use the extracted features predicted by this model as input.

iii)Sequence Processor. This is a word embedding layer for handling the text input, followed by a Long Short-Term Memory (LSTM) recurrent neural network layer.

iii)Decoder (for lack of a better name). Both the feature extractor and sequence processor output a fixed-length vector. These are merged together and processed by a Dense layer to make a final prediction.

The Photo Feature Extractor model expects input photo features to be a vector of 4,096 elements. These are processed by a Dense layer to produce a 256 element representation of the photo.The Sequence Processor model expects input sequences with a predefined length (10 words) which are fed into an Embedding layer that uses a mask to ignore padded values. This is followed by an LSTM layer with 256 memory units.Both the input models produce a 256 element vector. Further, both input models use regularization in the form of 50% dropout. This is to reduce overfitting the training dataset, as this model configuration learns very fast.The Decoder model merges the vectors from both input models using an addition operation. This is then fed to a Dense 256 neuron layer and then to a final output Dense layer that makes a softmax prediction over the entire output vocabulary for the next word in the sequence.

5)Generating New Descriptions:-

Once we have generated the model we can then use it for generating descriptions for new images.We have to see to it that the new images are similar to training images.We also have created a simple user interface that accepts any image in jpg/jpeg format and generates appropriate description.

V. TECHNOLOGY USED

1)Python:-

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead,when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

2)Keras:-

Keras is an open source neural network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or MXNet. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible.Keras contains numerous implementations of commonly used neural network building blocks such as layers, objectives, activation functions, optimizers, and a host of tools to make working with image and text data easier.Keras also provides tools to works with pre-trained neural so that we can change the layers of neural net as well as use it as stand-alone network for feature extraction.Keras doesn't require any modules other than Python.

VI. MODEL ARCHITECTURE

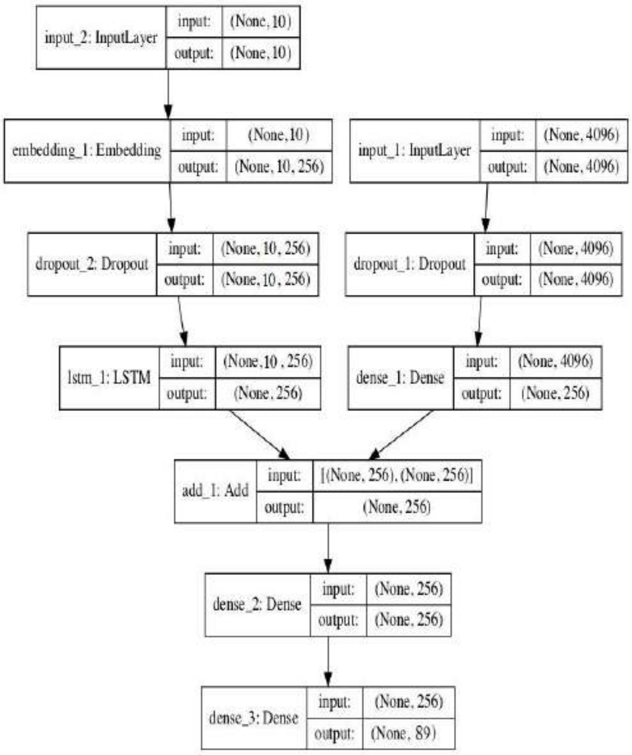


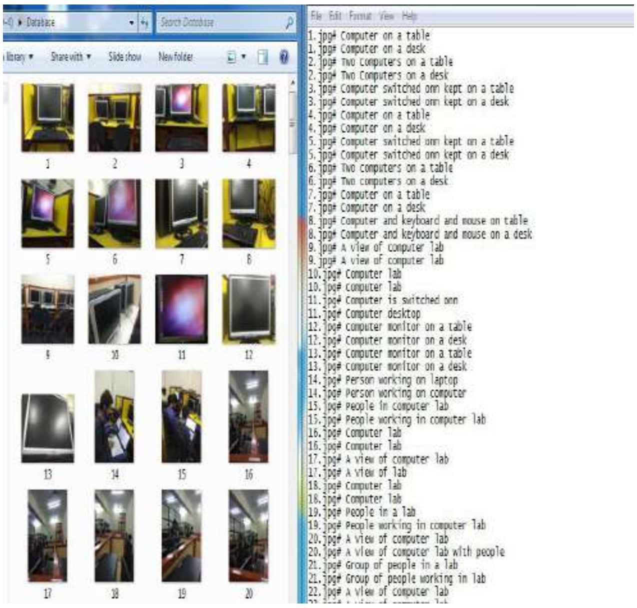
Figure 2: Summary of the combined Model for Image Description

VII. Experiment

We have given the description about the dataset we used and then we present our results.

a) Dataset

As in our model we are using a very small database we cannot find a database of such small size so we have to design our own dataset.While creating the Dataset we have taken care to include real time images of some day to day object and some images from the web.Also we have kept the ratio of images in training and validation(e.g.If there are 8 images of computer total we have to keep 6 images for Training,1 for Validation and 1 for Testing).Next we have described each image in few words for model to understand mapping of words to image features.We also have included two Description per image in order to have variety.



```
1.jpg# Computer on a table
1.jpg# Computer on a desk
2.jpg# Two computers on a table
2.jpg# Two Computers on a desk
3.jpg# Computer switched onn kept on a table
3.jpg# Computer switched onn kept on a desk
4.jpg# Computer on a table
4.jpg# Computer on a desk
5.jpg# Computer switched onn kept on a table
5.jpg# Computer switched onn kept on a desk
6.jpg# Two computers on a table
6.jpg# Two computers on a desk
7.jpg# Computer on a table
7.jpg# Computer on a desk
8.jpg# Computer and keyboard and mouse on table
8.jpg# Computer and keyboard and mouse on a desk
9.jpg# A view of computer lab
9.jpg# A view of computer lab
10.jpg# Computer lab
10.jpg# computer lab
11.jpg# Computer is switched onn
11.jpg# Computer desktop
12.jpg# computer monitor on a table
12.jpg# Computer monitor on a desk
13.jpg# Computer monitor on a table
13.jpg# computer monitor on a desk
14.jpg# Person working on laptop
14.jpg# Person working on computer
15.jpg# People in computer lab
15.jpg# People working in computer lab
16.jpg# Computer lab
16.jpg# Computer lab
17.jpg# A view of computer lab
17.jpg# A view of lab
18.jpg# Computer lab
18.jpg# Computer lab
19.jpg# People in a lab
19.jpg# People working in computer lab
20.jpg# A view of computer lab
20.jpg# A view of computer lab with people
21.jpg# Group of people in a lab
21.jpg# Group of people working in lab
22.jpg# A view of computer lab
```

Figure 3: Example Of the Dataset with Images and Descriptions

b) Performance Measurement

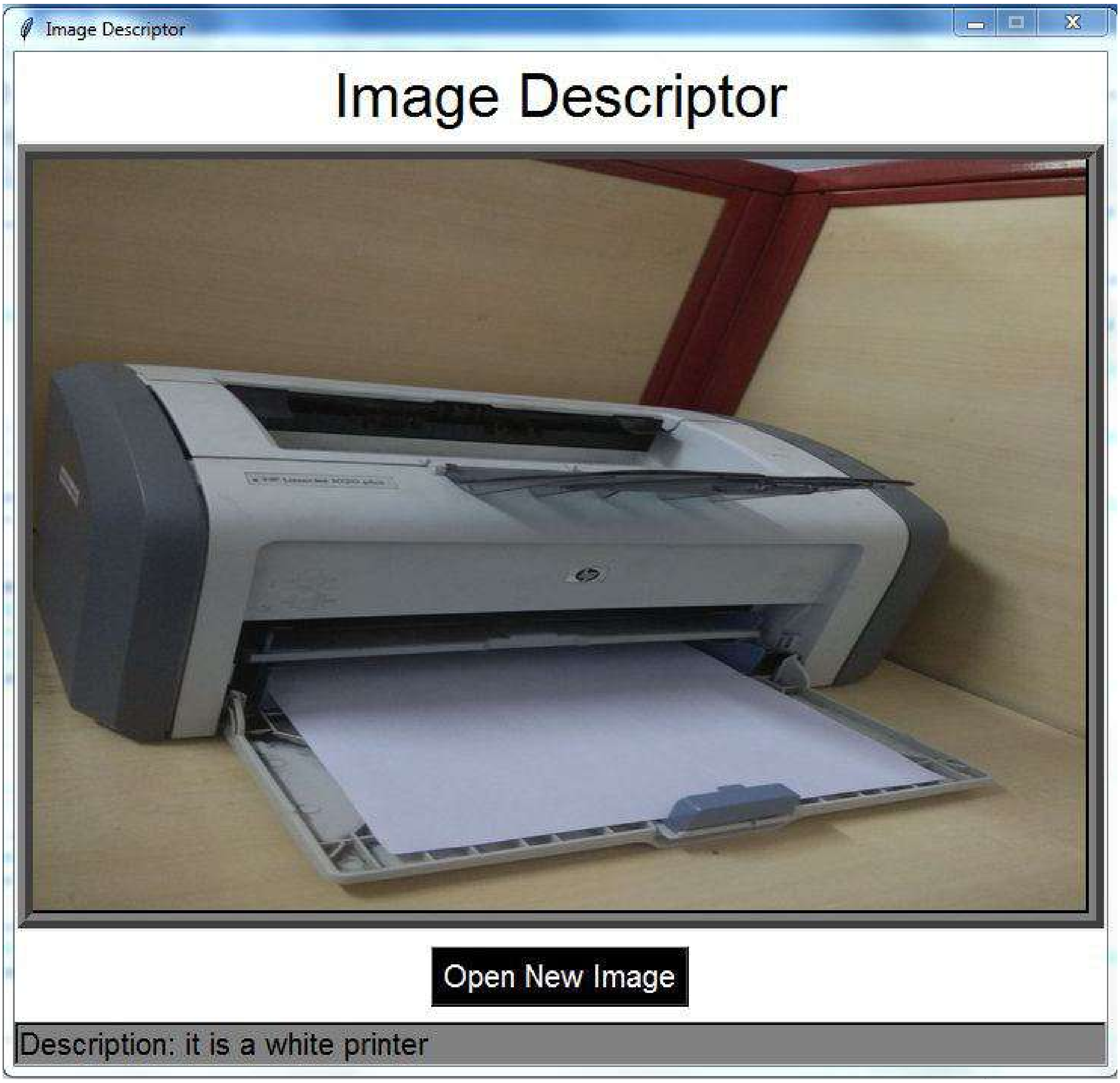
BLEU (bilingual evaluation understudy)[6] is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is" – this is the central idea behind BLEU.The BLEU score measures how many words overlap in a given translation when compared to a reference translation, giving higher scores to sequential words. BLEU scores range from 0-100, the higher the score, the more the translation correlates to a human translation.We will be using BLEU score to understand how well the model is able to generate proper descriptions.

c) Results

The proposed Image Description system has been implemented using Python 3.6 and requires Keras for manipulating the parts of Neural Network and uses Tensorflow in the Backend to do vector calculations.The Model takes any new image as input and can predict the description for that image in few words.We have also calculated BLEU scores for 25 test cases.We have plotted a graph that shows the variation in BLEU scores.

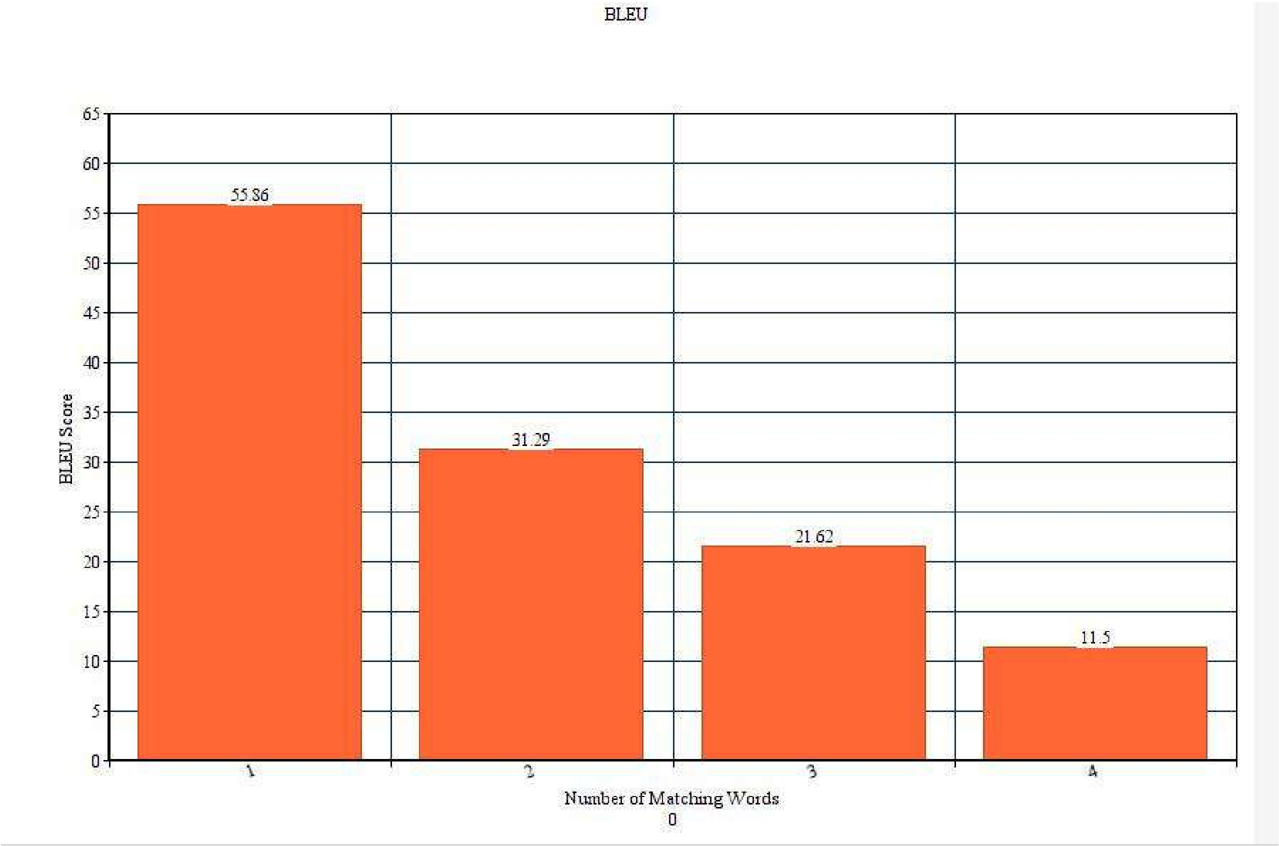
Figure 4 shows the interface for putting any image and shows description.

Graph 1 shows the graph of BLEU Scores



Description: it is a white printer

Figure 4: User Interface for the Model.



Graph 1: BLEU Scores for the Model

VII. CONCLUSION

Recent advances in artificial intelligence has provided us with ways to create such system through Neural Networks.Several other approaches are also mentioned but we have selected this method as it help us to learn a new emerging concept and offers the best solution.Our approach uses end to end system in which we can give input from one end and get output from other end.The use of CNN and RNN model in combination gives us this approach.We also have used every possible resource to reduce the overhead time generally required in Machine Learning Problems.We have also created our own database for applying the learned principles on a smaller scale.Our Model can generate description for images that are somewhat related to our training images.

IX. FUTURE SCOPE

This Model we have created is limited only to images and still requires a lot of dependency on external libraries and also requires Pre trained model to evaluate contents of images.In Future we can create such a model that has combined network of both the models and can work on video images i.e generating descriptions as the video is playing.

X. REFERENCES

[1]Karen Simonyan, Andrew Zisserman,”Very Deep Convolutional Networks for Large-Scale Image Recognition”, Computer Vision and Pattern Recognition(CVPR),2016

[2] Jiang Wang et al.,”CNN-RNN: A Unified Framework for Multi-Label Image Classification”The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[3] Alex Krizhevsky,Ilya Sutskever,Geoffrey E. Hinton,”ImageNet Classification with Deep Convolutional Neural Networks”,Advances in Neural Information Processing Systems 25(NIPS 2012)

[4] M.A.O. Vasilescu ; D. Terzopoulos,”Multilinear image analysis for facial recognition|”,Pattern Recognition, 2010. Proceedings. 16th International Conference.

[5] Tomas Mikolov et al.,”Recurrent neural network based language model.”

[6] BLEU (bilingual evaluation understudy) on Wikipedia

[7] Caption this, with Tensor Flow - O'Reilly Media(https://www.oreilly.com/learning/caption-this-with

-tensorflow).

[8]Google's Show and Tell: image captioning(https://research.googleblog.com/2016/09/show-and-tell-image-captioning-open.html).

Balance User Profile And Social Network Structure For Students And Faculty

Unmesh Patil[1] Sandesh Mankar[2] Sanjeet Rai[3] Abhijeet Pednekar[4]Samed Bhat[5]

[1,2,3,4,5] Department of Information Technology Engineering

[1,2,3,4,5]Pillai college of engineering,New Panvel,Maharastra,INDIA

Abstract-Social networks can enhance informal learning and support social connections within groups of learners and mentors. Social networks can help the development of communities. The Facebook platform provides an example of how a social networking service can be used as an environment for other tools connect, such as Chatbot for all student's general questions. The benefit of social networks is common interface which spans work as well as social boundaries. The existing methods focus on utilizing user and structure information alone. However, users information and structure information reflect different aspects of a user. The social networks that builds throughout one's existence networks that can consist of dozens, hundreds or even thousands of other people, with various degrees of mobility connectivity. We propose a network mapping method which integrates users and structural information. At first, the incorporative model represents user name, description, location information based on string matching, and friend information represented as relation network is regarded as structure information in social networking, with the development of social network technology, users often register accounts, post messages and create friends links on one platform. In social networks based on the proposed method, we develop a prototype system, which allows users to perceive various information. The experimental results on a real-world dataset demonstrate the efficiency of the proposed model with the intent of solving this problem that our work seeks to achieve success, as it analyzes the scientific social networks.

1 .INTRODUCTION

The project gives us unified platform for various services, such as social networking, storage, mail and messaging with simplifying access to other tools and applications that are provided over the internet which keeps the users up to date This can enhance informal learning and support social interactions within groups of learners and with those involved in the support of learning**.**

2 LITERATURE REVIEW

2.1 A Social Compute Cloud: Allocating and Sharing Infrastructure Resources via Social Network (2014), by Simon Caton, Christian Haas

Social network platforms have rapidly changed the way that people interact. They have enabled the establishment of, and participation in, digital communities as well as the representation, documentation and exploration of social relationships. We believe that as apps become more sophisticated, it will become easier for users to share their own services, resources and data via social networks.

2.2 Social network-based distributed data storage (2014) by Phani C. Polina , Bin Xie

Storing large amounts of data is challenging as it requires large reliable storage space. Currently, peer-to-peer (P2P) systems have been implemented for this purpose. However, these systems provide no guarantee of data retrieval as the data availability is determined by the interest of the users. On the other hand, cloud storage systems, built on top of a pool of powerful servers, can provide reliable data storage however, they are costly and vulnerable to privacy leakage.

2.3 Search in Social Networks (2014) by Ericsson Santana Marin , Cedric Luiz de Carvalho

In this paper, grounded in concepts from Network Science and Artificial Intelligence, we report on models we have constructed and on algorithms aimed at producing a search engine integrated into social networks environments. The contribution of this engine is its ability to evaluate the numerous paths that connect source and target people, opting for the path where the interpersonal influence through the path is maximized

2.4 A Social Group Utility Approach for Optimizing Computation Offloading in Cloudlet. (2016) by Ling Tang, Xu Che

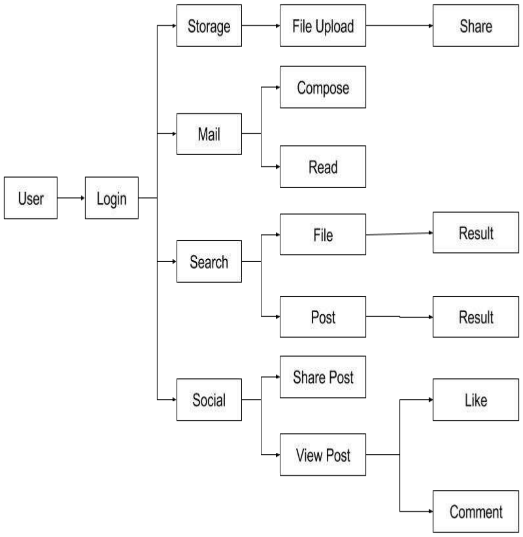
Cloudlet is a new paradigm in mobile cloud computing to provide resources to nearby mobile users via one-hop wireless connections. In this paper, we leverage the social tie structure among mobile users to achieve mutually beneficial computation offloading decision making, and hence, enhance the system-wide performance.

Proposed System

3.1.1 Existing System Architecture

There are several different online social networks, but some of them are most popular like – Facebook, LinkedIn etc. Each of these networks has its own unique style, functionality and patterns of usage. The first social networking site introduced in 1997.

3.1.2 Proposed System



3.2 Proposed System Architecture

Login : The process of identifying and authenticating themselves,by which an individual gains access to a system

. Users will be provided with the login ID and password, after registering to the websites.here the

MD5 algorithm is used for authentication for end users.After successful authentication to the website users will be able to access the website integrated facility services,
such as Storage, Mail, Search and Social.

INTEGRATED FACILITY SERVICES :

Storage: User can use this services for storing any file with different extension.
Storage is a file storage web service for storing and accessing data on forever connect infrastructure. The service combines the performance and scalability of internet with advanced security and sharing capabilities.

Mail :This enables the wide scope of connection across internet in order to establish a secure messaging service with addition of file or word. It has facility of sending and receiving the message in the form of composing mails and receiving mails. this mails can be filtered on basis of spams or different threats.

Search: Forever connect allows user to search different query such as a file with specific name or post of specific tagline.these result will displayed once the query is executed successfully.

Social: The function enables user to share an emotional feeling across the forever connect platform and also allows them to select the audition as per their choice.once the postis share it can be viewed by other user who are friend or connected on forever conect site.it also allow them to comment their feedback on that post. the last function enables them to like the post in favor of appreciation

Constraints of Project

- **Time:** The time constraint deals with the time necessary to finish a project. To successfully complete a project, the time constraint should be comprised of a schedule The estimated time for constructing the website is upto 6-8 months.
- **Cost:** Cost is another of the three constraints that you will want to become familiar with. The cost involved with successfully completing a project is dependent on a number of different elements, and some of these are material costs, the costs of labor, risk, and machines. The profit must also be analyzed when one is considering the cost constraint.Cost estimated in developing project is approximate ₹ 9,000 - 10,000.
- **Scope:** The third constraint of project management is scope. Scope can be defined as the tools and resources that are needed to achieve the end objective of the team. The scope can also be defined as the goal of the overall project, what it is supposed to achieve. Perhaps one of the most important aspects of the scope is the quality of the end product or service that is produced.
- **3.2 Implementation Details**
- The implementation details is given in this section.
-
- **3.2.1 Techniques:**
- **Password Security (Encryption):** A mathematical procedure for performing encryption on data. Through the use of an algorithm, information is made into meaningless cipher text and requires the use of a key to transform the data back into its

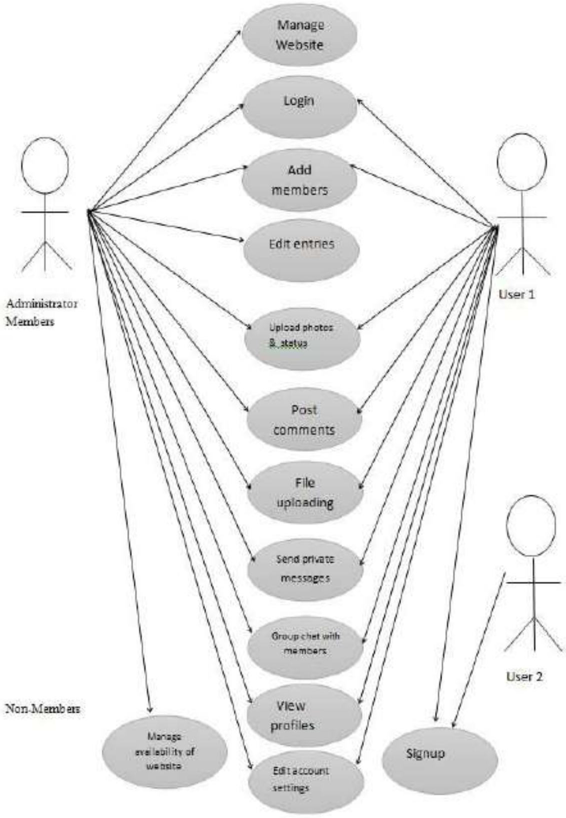
original form

Message-Digest algorithm 5(MD5: It is a hashing algorithm with cryptographic function and has been employed in a variety of security applications. It is used to check the integrity of files. It is typically expressed as a 32-digit hexadecimal number.

It processes a message of any length as input and returns a fixed-length output of 128 bits

Dijkstra's algorithm: It is an algorithm for finding the shortest paths between nodes. As a result, the shortest path algorithm is used in network routing protocols, most notably IS-IS (Intermediate System to Intermediate System) and Open Shortest Path First (OSPF). Dijkstra's algorithm also employed as a subroutine in other algorithms such as Johnson's

3.2.2 Use Case Diagram / Activity Diagram



3.3 Evaluation Metrics

LOC cost

Table 3.3: LOC cost

No.	Function or module	Estimated LOC
1	GUI	130
2	Database Operations	200
3	Backend	350
4	UI	170
	Total LOC	850

Complexity Factors Table

Table 3.4: Complexity Factors

F	Complexity Factors	Value Ratings out of 5
1	Are there distributed processing functions?	2
2	Is Performance Critical?	4
3	Is the code designed to be reusable ?	3
4	Will the system run in heavily utilized OS ?	3
5	Is the internal processing complex ?	4
		Total= 16

Cost Estimation

The basic COCOMO equations take the form

Software Project	a_b	b_b	c_b	d_b

Line of code(LOC)= 850 =0.85 KLOC

Effort Applied (E) = $a_b(KLOC)^{b_b}$ [man-months]

Effort =(2.4)(0.85)^1.05

Effort =2.11 Person-Months

Development Time (D) = $c_b(\text{Effort Applied})^{d_b}$ [months]

Development Time=2.5(2.11)^0.38

Development Time=1.88 Months

Applications

1) **Support for learning**

It support social connections within groups of learners and with those who involved in the support of learning.

2) **Support for members of an organization**
Social networks can be used by all members of an organization, and not just those involved in working with students. Social networks can help the development of communities of practice.

3) **Engaging with others**
It can provide valuable business intelligence and feedback on institutional services

4) **Ease of access to information and applications**
It can provide benefits to users by simplifying access to other tools and applications. Others social websites platform provides an example of how a social networking service can be used as an environment for other tools.

5) **Common interface**
A possible benefit of social networks may be the common interface which spans work as well as social boundaries. However, it can also be a barrier to those who wish to have strict boundaries between work and social activities.

Conclusion

This project is a novel distributed data storage system based on the social networks. Particularly, our system utilizes the social information of the users to find the potential storage nodes. The storage nodes are then selected based on the social ties with the data owner. In order to quickly obtain the potential storage nodes, we provide an efficient algorithm to search and compute the social ties in the social networks. The system performance is verified through both theoretical analysis and simulations. The results show that data stored in our proposed system is reliable and stable given the random nature of storage nodes joining and leaving the system

References

[1] Simon Caton, A Social Compute Cloud: Allocating and Sharing Infrastructure Resources via Social Networks, July-Sept. 2014.

[2] Ericsson Santana Marin, Search in Social Networks: Designing Models and Algorithms That Maximize Human Influence, 10 March 2014.

[3] Mohammad A. Salahuddin, A Survey on Content Placement Algorithms for Cloud-based Content Delivery Networks, 19 September 2017.

[4] Phani C. Polina, SOS: Social network-based distributed data storage, 10 March 2014.

[5] Liangmin Wang, A Cloud-Based Trust Management Framework for Vehicular Social Networks, 15 February 2017.

[6] Khaled Salah, Teaching Cybersecurity Using the Cloud, 20 April 2015.

[7] Changqin Huang, On Selecting Vehicles as Recommenders for Vehicular Social Networks ,2017

[8] Seeing Is Believing: Sharing Real-Time Visual Traffic Information via Vehicular Clouds,Liviu Iftode,2016.

Comparative analysis of Botnet IDS based on classification and clustering techniques

Prof. Payel Thakur
Department of Computers
Pillai College of Engineering
Panvel, India

Jatin Rajan
Department of Information Technology
Pillai College of Engineering
Panvel, India

Manish Poojari
Department of Information Technology
Pillai College of Engineering
Panvel, India

Nitesh Jha
Department of Information Technology
Pillai College of Engineering
Panvel, India

Karan Nair
Department of Information Technology
Pillai College of Engineering
Panvel, India

***Abstract*—Botnet detection plays an important role in network security. For detecting the presence of bots in any network, there are many detection techniques available. Intrusion Detection Systems (IDS) functions as an efficient counter against botnets. However, there are several methods to implement an IDS based on the configuration and various requirements of the network. Data Mining is one the analysis methods used by many IDSs' for recognizing an attack from botnets. The process of data mining can involve many techniques such as classification methodology, which implements well known classification schemes or data partitioning via clustering algorithms. We propose an implementation of Average One Dependence Estimators (A1DE)[1] which is a recent enhancement of naive bayes classification algorithm and Maximum density clustering and T-IDS[2] which is built on randomized data partitioned learning model (RDPLM) and evaluate it's performance in real time. Training and Testing of the RDPLM system is done using NSL-KDD data set. We will then do a comparitive study on these botnet detection techniques.**

Keywords—Average One Dependence Estimators (AODE); Botnet; Naïve Baye; Traffic based IDS

I. Introduction

Botnet is a network of some infected computers (bots), these bots are controlled by botmaster through command and control (C&C) activities. It is a logical collection of internet connected devices such computers, smartphones or IoT devices whose security has been breached by software from a *malware* (malicious software) distribution and control is transferred to a third party. The controller of a botnet is able to direct the activities of these compromised computers through communication channels formed by standards-based network protocols such as IRC and Hypertext Transfer Protocol(HTTP)[3] or the owner can control the botnet using command and control (C&C) software. The botnets cause spam, DDoS and some high-damaged attacks, steal data,[4] allow the attacker access to the device and its connection, and affect to almost organizations or personals.

Botnet Intrusion detection Systems focus on detecting whether the network is under the attack of a botnet or can be used to analyse a system and see if it's affected by a botnet. The study and innovations on Botnet is vast since many techniques are being developed to detect myriads botnets which are being used by attackers without devoting too much processing power or CPU time to the process.

II. Literature Survey

A. Comparison of Clustering Algorithms (Long Mai et al., 2016)

Comparison of detection rate is done for 3 algorithms: K-means, DBSCAN and Mean Shift clustering and combine with Decision Tree classification to create hybrid model. Mean Shift clustering gets better performance in comparison to other to algorithm. But when we increase training data size, K-means outperforms other algorithms.

B. *Botnet Detection based on Anomaly and Community Detection (Jing Wang et al., 2017)*

Proposes a novel two-stage approach for detecting Botnets.

Stage1: The first stage detects anomalies by leveraging large deviations of an empirical distribution.

Stage2: The second stage identifies the bots by analyzing these anomalies using pivotal nodes (botmaster or target) that interact with bots.To characterize this correlation, they construct a Social Correlation Graph. Since bots and behaviours are correlated, they are mostly connected to each other. Then the appropriate division of the SCG to separate bots and normal nodes is found to make the system most effective.

C. *Botnet Recognition Method Based on Fuzzy Classification (Dong Wang et al., 2016)*

They use fuzzy clustering, anomaly detection, fuzzy pattern recognition and fuzzy association rules to identify botnet. The first step: According to their data collection set (real zombie network traffic) they carry out a detailed analysis and also extraction of UDP and TCP traffic in features. The second step: the characteristics of the data set of fuzzy clustering and division level are set. The third step is the use of fuzzy recognition level, which is calculated for each feature support, trust, and finally find out the association rules which will be used for botnet detection.

D. *Botnet Identification Via Universal Anomaly Detection (Shachar Siboni et al., 2014)*

The IDS is built on a universal anomaly detection system, which is protocol independent, and is applied to detect C&C channels and botnets in data set with a negligible false alarm probability.

E. *Query or Spams: Detecting fraudulent web requests using stream clustering (Tahere Shaqiba et al., 2015)*

In this paper they propose a method based on a semi-supervised stream clustering algorithm which analyzes the activity. And K fold cross validation which resulted in a satisfactory accuracy.

F. *DGA Botnet Detection Utilizing Social Network Analysis (Tzy-Shiah et al., 2016)*

This study proposes a DGA botnet detection mechanism utilizing the feature-based characteristics of social networks. Domain Generation Algorithm (DGA). All bots in a DGA botnet periodically execute a DGA to generate a list of candidate C&C domains. Each bot then performs DNS queries for the domains in the list one by one until it connects to the C&C server. When a domain is detected and blocked by defending systems, the botmaster simply migrates the C&C server by associating it with a new IP and a new domain name in the list of candidate C&C domains. The results show that the proposed mechanism has the ability to accurately and effectively detect both well-known and new malicious DGA botnets in real-world networks.

G. *Exploring a Service-Based Normal Behaviour Profiling System for Botnet Detection (Xiao Luo et al., 2017)*

Profiling-based botnet detection using three learning algorithm self-organizing map (SOM), local outlier factor (LOF), and k-NN outlier factor. SOM is a type of artificial neural network (ANN) that is trained using unsupervised learning to generally produce a two-dimensional, discretized representation of the input space of the training samples, called a map,The local outlier factor is based on a concept of a local density (like DBSCAN and OPTICS), where locality is given by k nearest neighbors, whose distance is used to estimate the density. The system has 91 % detection rate with a false alarm rate of 5 %.

H. *Modified K-Mean Algorithm Using Timestamp Initialization in Sliding Window to Detect Anomaly Traffic (Putra et al., 2015)*

The designed Botnet IDS implements K-Means algorithm using Timestamp Initialization in which where the cluster initialization was used(Simple K-Means Algorithm) Timestamp Initialization as applied. Expected modified K-Means using Timestamp Initialization can eliminate the determination of K-cluster that affect detection rate and false positive rate when using different K-cluster. Hence IDS obtains a high detection rate and low false positive rate, Testing was done via KDDCup'99 dataset which is also used for training.

I. *Peer to Peer Botnet Detection Based on Network Traffic Analysis (Suzan Almutairi et al., 2016)*

The paper addresses the problem of detection P2P botnets by creating a IDS using a network analyzer which monitors two activities:

Activity 1: Bot connection and bot communication with other bots or with the C&C server.

Activity 2: Network data stream.

Done via preprocessing the data and detect the bots based on specific rules. The algorithm extracts the feature values for each P2P in the network and preprocesses them.

A detection report is provided which is responsible for producing the final score result based on the network analysis and provides a report of infected machines. The results of experiments show a high level of accuracy

(99.1%) and a low positive rate (0.04).

III. Existing System Architecture

The Intrusion Detection Systems(IDS) need proper techniques for analysing network data, and some strategies for deciding whether the system is under attack via intrusion carried out by Bots under the control of a botmaster. Existing Botnet IDS include Signature based IDS and Anomaly based IDS.

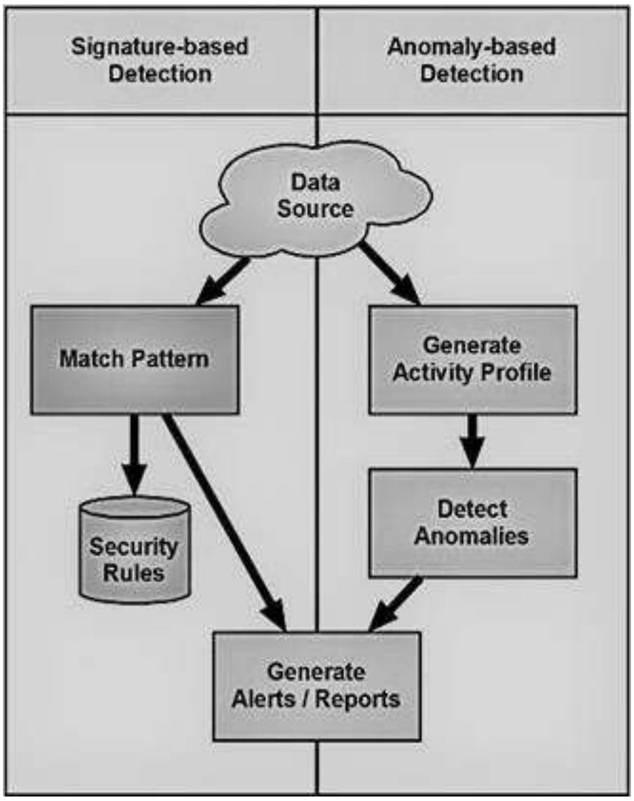


Fig. 1. Existing system architecture used for Both Signature and Anomaly based IDS[5]

Signature Based IDS**:** This form of IDS is quite similar to Virus Scanner. It analyses the entire network data and searches for known identities or signatures for any specific intrusion event. Signature based IDS is very efficient at detecting known attacks as it uses little computing resources and time. However Signature based IDS are only as good as the signatures stored in it's database. Hence it relies heavily on timely updates to keep in touch with variations of hacker techniques. Even then it is quite possible to generate false positive on SNORT IDS[6] and has been empirically tested for signature based IDS[7].

Anomaly Based IDS: Signature and anomaly-based systems are similar in terms of conceptual operation and composition. Signature-based schemes provide very good detection results for specified, well-known attacks. However, they are not capable of detecting new, unfamiliar intrusions. The main benefit of anomaly-based detection techniques is their potential to detect previously unseen intrusion events. The normal (or abnormal) behaviour of the system is characterized and a corresponding model is built. Once the model for the system is available, it is compared with the observed traffic at the host system or network in question. If the deviation found exceeds (or is below, in the case of abnormality models)[8] a given threshold an alarm will be triggered. Even if anomaly based IDS covers unfamiliar intrusions, false positive rates of anomaly are greater than that of signature based IDS.[9]

IV. Proposed System Architecture

The previous sections discussed the strengths and weaknesses of existing system. In order to achieve better results for detection under the assumption of decent hardware capabilities we have proposed a dual running IDS that parallelly runs Classification and Clustering variants which provide their results after computation to a decision making module. The proposed architecture of IDS application is shown in Figure 3.3

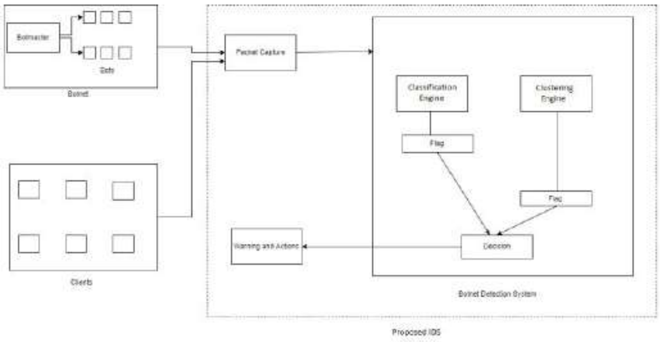


Fig. 2. Proposed system architecture

Data Packets coming from all clients will first be captured using Packet Capture tool (either Wireshark 2.4.2 or if time permits hard coded packet capture tool integrated to the IDS Application) and once the packets are captured, necessary parameters and data is sent to our core application.

The core application will run both classification and clustering engines (A1DE and Max Density clustering)parallelly. Once the computation results are obtained, they are sent to the Decision making module. The module will decide based on pre-decided parameters whether to generate alerts. The alerts will be visible on the custom made GUI for our IDS application.

The custom made GUI will be used to provide alerts to the user of the Host System/Server. Separate computational results for both Classification and Clustering engines can also be viewed using the GUI. GUI will also provide with a comparitive analysis of both engines if we simulate an attack on a virtual environment.

A. Implementation Details

The entire process of botnet attack on a victim system will be done in a simulated environment. Network simulation will be done using Wireshark which will run in tandem with Virtual Box. Multiple nodes will be created which will carry out of the following roles:

Botmaster**:** The System used to orchestrate the attack.

Bots**:** Zombie systems which will execute the attack.

Victim System: The system which will contain our Application IDS.

Legitimate Clients: Simple clients accessing Victim system for non-malicious purposes.

The network can be changed in accordance to our needs or analysis purposes using images of Routers and Switches to cascade many networks (Also adapters can be simulated in Virtual Box for direct node connection)

B. Algorithms/Methodology/Techniques

Simulation will involve sending data packets from legitimate clients and bots. Further specifications involved in the simulation are as follows:

- Data packet capture:

Done by using Wireshark or a separate coded packet capture tool. Once packets are captured they will be sent for further analysis.

- Current Methodology (Future possibility of an IDS application):

Packets are obtained from the Victim PC using Wireshark in Capture Mode. The contents obtained via Wireshark are copied into a text file which is later imported to Microsoft Excel. Using Excel, the document is converted into “.csu” extension which is necessary for WEKA application set documents.

Once opened in WEKA (under .arss extension) file will contain parameters such Time, Source, Length but not the information within the packets. After which the data is analyzed based on the parameter of length. If the value is too abnormal it is considered as an outlier with suspicious activity being carried out. Analysis is done using the

following classification and clustering engines:

1.A1DE(Average One Dependence Estimators)

2.Naïve Bayes

3.K-Means

4.Max Density Clustering

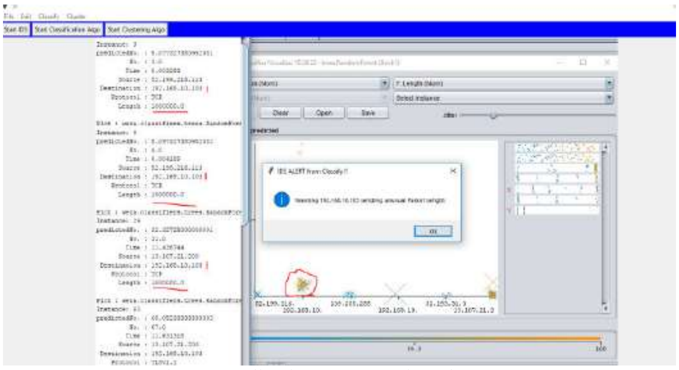
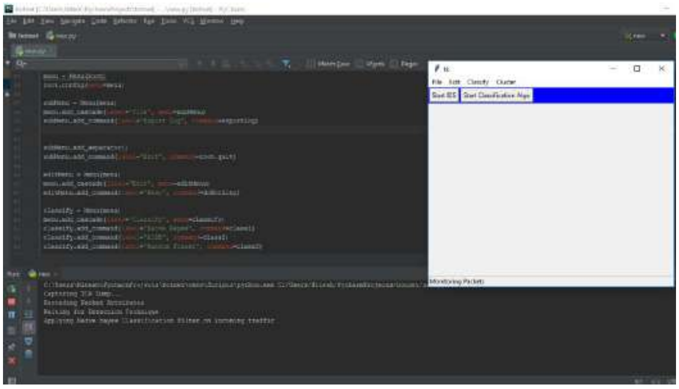

File Edit Classify Cluster
Start IDS
Start Classification Algo
Start Clustering Algo
Clear
Open
Save
predicted
IDS ALERT from Classify !!
OK

Fig. 3. Application GUI creating alerts

An IDS Application will be created either using Neutron IDE or NetBeans IDE(for Java) This application will involve classification and clustering computation engines and GUI which the user will interact with.The IDS application involves only threat alerts and computation reports. The application can be further enabled to carry out actions such as “Block” or ”Track” which will require additional code blocks and their integration the original application.



```
Capturing TCP Dump...
Extracting Packet Attributes
Waiting for Detection Technique
Applying Naive Bayes Classification Filter on incoming traffic
```

Fig. 4. Application GUI created for classification and clustering engine

V. Summary

In this report, the study of different IDS is presented. The different techniques such as Classification, Clustering, Signature based, and Anomaly based are analysed. On the basis of these techniques selected from myriads of techniques stated in literature survey, a parallel application IDS was proposed. The application provides a comparitive analysis of two state of the art and new classification and clustering variants in a virtualized environment. This virtualized environment created via network simulation provides a platform for studying IDS techniques and analysing them without causing harm to the system by creating our own Botnet and Victim System. The parallel approach is proposed with a future scope of “Action” module. The performance measures like false positives rate and detection rate are described in this report. The applications of this domain is also identified.

Acknowledgment

We like to thank our Mentor Prof. Payel Thakur without whom this project would not have been possible. We would also like to thank our HOD Dr. Sharvari Govilkar for her continued support and guidance.

We would also like to extend our gratitude to our Principal Dr. R. I. K. Moorthy for providing us this opportunity. Also a special thanks to all teaching and non teaching staff for providing us with the equipments and environment necessary for this project.

References

[1] Amreen Sultana, M. A. Jabbar; “Intelligent network intrusion detection system using data mining techniques”, 2016 IEEE.

[2] Omar Y. Al-Jarrah, Omar Alhussein, Paul D. Yoo, Sami Muhaidat, Kamal Taha, Kwangjo Kim; “Data Randomization and Cluster-Based Partitioning for Botnet Intrusion Detection”, 2016 IEEE

[3] Ramneek, Puri; "Bots & Botnet: An Overview", SANS Institute, 2003.

[4] "Thingbots: The Future of Botnets in the Internet of Things", 20 February 2016.

[5] <https://www.researchgate.net/figure/297171228_fig1_FIGURE-1-SIGNATURE-AND-ANOMALY-BASED-IDS-5>

[6] SNORT [written by Martin Roesch] <http://www.snort.org/>

[7] Samuel Patton, William Yurcik, David Doss; “An Achilles’ Heel in Signature-Based IDS: Squealing False Positives in SNORT”, Illinois state university USA 2008

[8] Este´vez-Tapiador JM, Garcı´a-Teodoro P, D´ıaz-Verdejo JE; “Anomaly detection methods in wired networks: a survey and taxonomy”, Computer Networks 2004;27(16):1569–84.

[9] Garcia Teodoro, diaz verdejo, G. Marcia Fernandez, e Vazquez, “Anomaly based intrusion detection” ,2009 <www.elsevier.com/locate/cos>

Best Keyword Cover Search

Aditya Lekhak, Priya Bisht, Sukanti Bandabe, Siddhesh Deshpande

Department Of Information Technology

Pillai College of Engineering, New Panvel - 410 206

University Of Mumbai

Academic Year 2017-2018

Abstract-Objects in a spatial database like hotels and restaurant are associated with keywords to indicate their business and features. A problem called as Closest Keywords search is to query keyword Cover (an object), which checks a set of query keywords and give the minimum inter-objects distance. Nowadays we observe an increase in availability and importance of keyword rating in object evaluation, this helps in better decision making and motivates us to check a version of Closest Keywords search called Best Keyword Cover. This method consists of inter-objects distance along with the keyword rating of objects. Baseline algorithm combines objects from different search keywords to generate candidate keyword covers. The performance of baseline algorithm drops when query keywords increases, as a result of massive candidate keyword covers generated. To counter this difficulty, a scalable algorithm called keyword nearest neighbor expansion (k-NNE) is used.

k-NNE significantly reduces the number of candidate keyword covers generated, the in-depth analysis has proven the supremacy of k-NNE algorithm.

Index Terms—Spatial database, Point of Interests, Keywords, Keyword Rating, Keyword Cover

INTRODUCTION

Driven by mobile computing, location-based services and wide availability of comprehensive digital maps and satellite imagery (e.g., Google Maps and Microsoft Virtual Earth services), the spatial keywords search problem has attracted much attention recently [2,3].

In a spatial database, each tuple represents a spatial object which is associated with keyword(s) to indicate the information such as its businesses/services/features. The essential task of spatial keywords search is to identify spatial object(s) which are associated with keywords applicable to a set of query keywords, and have good spatial relationships (e.g., close to each other and/or close to a query location). This problem has unique value in various applications because users' requirements are often expressed as multiple keywords. For example, a tourist who plans to visit a city may have particular shopping, dining and accommodation needs. It is desirable that all these needs be satisfied without long distance travelling.

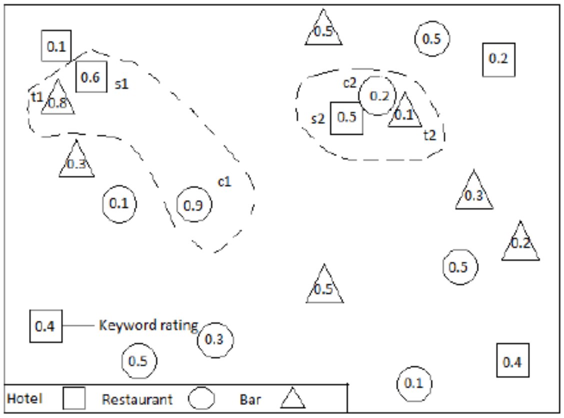


Fig. 1. BKC vs. mCK

The m-closest keywords (mCK) [6,7] query searches for a group of objects that covers all query keywords close to each other. The common version of mCK query is Best Keyword Cover (BKC) query. It considers inter-objects distance as well as keyword rating. This is used for increasing availability and importance of keyword rating for better decision making.

Consider example as shown in figure 1. Suppose the query keywords are "Hotel", "Restaurant" and "Bar". mCK query returns $\{t_2, s_2, c_2\}$ since it considers the distance between the returned objects only. BKC query returns {t₁, s₁, c₁} since the keyword ratings of object are considered in addition to the inter-objects distance. Compared to mCK query, BKC query supports more robust object evaluation and thus underpins the better decision making.

Baseline and keyword-NNE are BKC query processing algorithms. The baseline algorithm is motivated by the mCK query processing methods. Both the baseline algorithm and keyword-NNE algorithm are supported by indexing the objects with an R*-tree like index, called $\mathrm{KRR}^* - \mathrm{tree}$[1]. The baseline algorithm, combines nodes in higher hierarchical levels of KRR*-trees. This will generate candidate keyword covers. Then, the most promising candidate is assessed in priority by combining their child nodes to generate new candidates. Even though BKC query can be effectively resolved, when the number of query keywords increases, the performance drops dramatically as a result of massive candidate keyword covers generated.

To overcome this critical drawback, keyword nearest neighbor expansion (keyword-NNE) algorithm is used which applies a different strategy. Keyword-NNE selects one query keyword as principal query keyword. The objects associated with the principal query keyword are principal objects. For each principal object, the local best solution (known as local best keyword cover (lbkc)) is computed. Among them, the lbkc with the highest estimation is the solution of BKC query. Given a principal object, its lbkc can be identified by simply retrieving a few nearby and highly rated objects in each non-principal query keyword. Compared to the baseline algorithm, the number of candidate keyword covers generated in keyword-NNE algorithm is significantly reduced. The in-depth analysis reveals that the number of candidate keyword covers further processed in keyword-NNE algorithm is optimal, and each keyword candidate cover processing generates much less new candidate keyword covers than that in the baseline algorithm.

PRELIMINARY

In a spatial database, objects are associated with one or many keywords. Without loss in principle, the object with many keywords are transformed to multiple objects located near-by. So, the object is in the form <id, x, y, keyword, rating> where x, y defines the location of the object in a 2-dimensional space.

Definition 1 (Diameter): Let O be a set of objects $\{\mathrm{o}_1, \dots, \mathrm{o}_\mathrm{n}\}$. For o_i, o_j $\in$ O, $\mathrm{dist}(\mathrm{o}_\mathrm{i}, \mathrm{o}_\mathrm{j})$is the Euclidean distance between o_i, o_j. The diameter is given by [1]:

$$\mathrm{diam}(\mathrm{O}) = \max_{\mathrm{o}_\mathrm{i},\mathrm{o}_\mathrm{j} \epsilon \mathrm{O}} \mathrm{dist}(\mathrm{o}_\mathrm{i}, \mathrm{o}_\mathrm{j}) \qquad (1)$$

The score of O is a function with respect to diameter of and keyword rating of objects. Users have varied interests in keyword rating of objects. The linear interpolation function [3,5] is used to obtain the score of O.

$$\begin{aligned} \mathrm{O.score} &= \mathrm{score}(\mathrm{A}, \mathrm{B}) \\ &= \alpha\left(1 - \frac{\mathrm{A}}{\mathrm{max\_dist}}\right) + (1-\alpha)\frac{\mathrm{B}}{\mathrm{max\_rating}} \\ \mathrm{A} &= \mathrm{diam}(\mathrm{O}) \qquad (2) \\ \mathrm{B} &= \mathrm{min}_{\mathrm{o}\in\mathrm{O}}(\mathrm{o.rating}) \end{aligned}$$

where B is the minimum keyword rating of objects in O and α (0 ≤ α ≤ 1) is an application specific parameter. If α = 1, the score of O is determined only by the diameter of O. If α = 0, the score of O only considers the minimum keyword rating of objects in Q where $\mathrm{max\_dist}$ and $\mathrm{max\_rating}$ are used to normalize diameter and keyword rating into [0,1] respectively. $\mathrm{max\_dist}$ is the maximum distance between any two objects in the spatial database D, and $\mathrm{max\_rating}$ is the maximum keyword rating of objects.

Proposition 1: The score is of uniform property [1].
Proof: Given a set of objects O_i, where O_j is a subset of O_i . The diameter of O_i should not be less than of O_j , and the minimum keyword rating of objects in O_i must be not greater than that of objects in O_j . Therefore,$\mathrm{O}_\mathrm{i}.\mathrm{score} \leq \mathrm{O}_\mathrm{j}.\mathrm{score}$.

Proposition 2: Given two keyword covers O and O', O' consists of objects $\{\mathrm{o}_{\mathrm{k}1}, \dots, \mathrm{o}_{\mathrm{kn}}\}$ and O consists of nodes $\{\mathrm{N}_{\mathrm{k}1}, \dots, \mathrm{N}_{\mathrm{kn}}\}$. If o_{ki} is under N_{ki} in $\mathrm{KRR}^*{}_\mathrm{k} - \mathrm{tree}$ for $1 \leq \mathrm{i} \leq \mathrm{n}$, it is true that O'.score ≤ $\mathrm{O.score}$[1].

Definition 2: Let T be a set of keywords {k1, · · ·,kn} and O be a set of objects {o1, · · ·, on}, O is the keyword cover of T if one object in O is associated with one and only one keyword in T.

Definition 3: In a spatial database D and a set of query keywords T, BKC query returns a keyword cover O of T $\mathrm{O} \subset \mathrm{D}$ such that O.score $\geq$ $\mathrm{O}'.\mathrm{score}$ for any keyword cover O' of T ($\mathrm{O}' \subset \mathrm{D}$).

Notations used in this paper are summarized below [1]:

Notation	Interpretation
D	A spatial database
T	A set of query keywords
O_k	The set of objects associated with keyword k.
o_k	An object in O_k.
KC_o	The set of keyword covers in each of which o is a member.
kc_o	A keyword cover in KC_o.
$lbkc_o$	The local best keyword cover of *o, i.e.,* the keyword covers in KC_o with the hightest score.
$o_k.NN_{ki}^n$	o_k 's nth keyword nearest neighbor in query keyword k$_i$.
$KRR^*{}_k - tree$	The keyword rating R*-tree of O_k.
N_k	A node of KRR*$_k$-tree.

TABLE 1
SUMMARY OF NOTATIONS:

INDEXING KEYWORD RATINGS

To process BKC query, we use R*-tree with an additional dimension for keyword rating. Here, a 3-dimensional R*-tree called keyword rating R*-tree (KRR*-tree) [1] is used. The range of both spatial and keyword rating dimensions are normalized to [0,1]. Suppose we need construct a KRR*-tree over a set of objects D. Each object o ∈ D is mapped into a new space using the following mapping function [1]:

$$\mathrm{f{:}\,o(x,y,rating)} \rightarrow \left(\frac{\mathrm{x}}{\mathrm{max_x}}, \frac{\mathrm{y}}{\mathrm{max_y}}, \frac{\mathrm{rating}}{\mathrm{max\_rating}}\right). \qquad (3)$$

where $\mathrm{max_x}$, $\mathrm{max_y}$ and $\mathrm{max_y}$ are the maximum value of objects in D on x, y and keyword rating dimensions respectively. Each node N in KRR*-tree is defined as $\mathrm{N(x,y,r,l_x,l_y,l_r)}$ where x is the value of N in x axle close to the origin, and $\mathrm{l_x}$ is the width of N in x axle, so does y, $\mathrm{l_y}$ and r, $\mathrm{l_r}$.

BASELINE ALGORITHM

The baseline algorithm is inspired by the mCK query processing methods . For mCK query processing, the method in browses index in top-down manner while the method in does bottom-up. Given the same hierarchical index structure, the top-down browsing manner The baseline algorithm is inspired by the mCK query processing methods . For mCK query processing, the method in browses index in top-down manner while the method in does bottom-up. Given the same hierarchical index structure, the top-down browsing manner The baseline algorithm is inspired by the mCK query processing methods . For mCK query processing, the method in browses index in top-down manner while the method in does bottom-up. Given the same hierarchical index structure, the top-down browsing manner is used. When designing the baseline algorithm for BKC query processing we take advantage of both the methods. First we apply multiple KRR-trees which contain no keywords information in nodes such that the number of nodes of the index is not more than of the index in second, the top-down index browsing method can be applied since each keyword has own index.

Algorithm Baseline(T:Root)
Input: A set of query keywords T the root nodes of all KRR*-trees Root.

Output- Best Keyword Cover.

```
1   BKC;
2   H Generate Candidate(T; Root; bkc);
3   While H is not empty do
4   can the candidate in H with the highest score;
5   Remove can from H;
6   Depth First Tree Browsing(H; T; can; bkc);
7   for each candidate 2 H do
8   if(candidate : score bkc: score) then
9   remove candidate from H;
10  return bkc;
```

Algorithm shows the pseudo-code of the baseline algorithm. Given a set of query keywords T, it first generates candidate keyword covers using Generate Candidate function which combines the child nodes of the roots of KRR*$_{ki}$-trees for all k$_i$ 2 T (line 2). These candidates are maintained in a heap H. Then, the candidate with the highest score in H is selected and its child nodes are combined using Generate Candidate function to generate more candidates. Since the number of candidates can be very large, the depth-first KRR*$_{ki}$-tree browsing strategy is applied to access the leaf nodes as soon as possible (line 6). The first candidate consisting of objects (not nodes of KRR*-tree) is the current best solution, denoted as bkc, which is an intermediate solution.

KEYWORD NEAREST NEIGHBOR EXPANSION (KEYWORD-NNE)

Using the baseline algorithm, BKC query can be effectively resolved. However, it is based on exhaustively combining objects (or their MBRs).

Even though pruning techniques have been explored, it has been observed that the performance drops dramatically, when the number of query keywords increases, because of the fast increase of candidate keyword covers generated. This motivates us to develop a different algorithm called keyword nearest neighbor expansion (keyword-NNE). We focus on a particular query keyword, called principal query keyword. The objects associated with the principal query keyword are called principal objects. Let k be the principal query keyword. The set of principle objects is denoted as Ok.

Definition 4 (Local Best Keyword Cover): Given a set of query keywords T and the main query keyword $k \in T$, the local best keyword cover of a main object o_k is

$$\mathrm{lbkc}_{ok} = \left\{\mathrm{kc}_{ok} \middle| \mathrm{kc}_{ok} \epsilon \mathrm{KC}_{ok}, \mathrm{kc}_{ok}.\mathrm{score} = \max_{\mathrm{kc} \epsilon \mathrm{KC}_{ok}} \mathrm{kc.score}\right\}. \quad (4)$$

Where KC_{ok} is the set of keyword covers in each of which the principal object o_k is a member.

For each principal object $o_k \ \epsilon \ O_k$, lbkc_{ok} is identified. Among all principal objects, the lbkc_{ok} with the highest score is called global best keyword cover (GBKC_k).

Proposition 3: GBKC_k is the solution of BKC query. Proof: Assume the solution of BKC query is a keyword cover kc other than GBKC_k, i.e., $\mathrm{kc.score} > \mathrm{GBKC}_k.\mathrm{score}$. Let o_k be the principal object in kc.By definition, $\mathrm{lbkc}_{ok}.\mathrm{score} \geq \mathrm{kc.score}$, and $\mathrm{GBKC}_k.\mathrm{score} \geq \mathrm{lbkc}_{ok}.\mathrm{score}$. So, $\mathrm{GBKC}_k.\mathrm{score} \geq \mathrm{kc.score}$ must be true. This conflicts to the assumption that BKC is a keyword cover kc other than GBKC_k.

Step 1. One query keyword $k \in T$ is selected as the principal query keyword;
Step 2. For each principal object $o_k \in O_k$, lbkc_{ok} is computed;
Step 3. In O_k, GBKC_k is identified;
Step 4. return GBKC_k.

LBKC Computation

Given a principal object o_k, lbkc_{ok} consists of ok and the objects in each non-principal query keyword which is close to ok and have high keyword ratings. It motivates us to compute lbkc_{ok} by incrementally retrieving the keyword nearest neighbors of ok.

Keyword Nearest Neighbor-

Definition 5 (Keyword Nearest Neighbor (Keyword-NN)):
Given a set of query keywords T , the principal query keyword is $k \in T$ and a non-principal query keyword is $k_i \in T/\{k\}$. O_k is the set of principal objects and O_{ki} is the set of objects of keyword k_i. The keyword nearest neighbor of a principal object $o_k \in O_k$ in keyword k_i is $o_{ki} \in O_{ki}$ if $\{o_k, o_{ki}\}.\mathrm{score} \geq \{o_k, o'_{ki}\}.\mathrm{score}$ for all $o'_{ki} \in O_{ki}$.

The first keyword-NN of o_k in keyword k_i is denoted as $o_k.\mathrm{nn}_{ki}^1$ and the second keyword-NN is $o_k.\mathrm{nn}_{ki}^2$, and so on. These keyword-NNs can be redeem by browsing KRR*ki-tree. Let N_{ki}be a node in KRR*ki-tree. $\{o_k, N_{ki}\}.\mathrm{score} = \mathrm{score}(A, B)$. (5)

$$A = \mathrm{dist}(o_k, N_{ki}).$$
$$B = \min(N_{ki}.\mathrm{maxrating}, o_k.\mathrm{rating}).$$

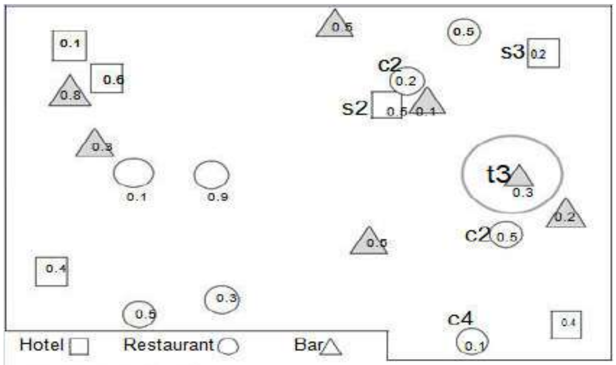
where $\mathrm{dist}(o_k, N_{ki})$ is the minimum distance between o_k and N_{ki} in the 2-dimensional geographical space defined by x and y dimensions, and $N_{ki}.\mathrm{maxrating}$ is the maximum value of N_{ki} in keyword rating dimension.

Preposition 4: For any object o_{ki} under node N_{ki} KRR*ki-tree,

$$\{o_k, N_{ki}\}.\mathrm{score} \geq \{o_k, o_{ki}\}.\mathrm{score} \quad (6)$$

Proof: It is a special case of Preposition 2.

To retrieve keyword-NNs of a principal object o_k in keyword ki, KRR*ki-tree is browsed in the best-first strategy [9]. The root node of KRR*ki-tree is visited first by keeping its child nodes in a heap H. For each node $N_{ki} \in H, \{o_k, N_{ki}\}.\mathrm{score}$ is computed. The node in H with the highest score is replaced by its child nodes. This operation is repeated until an object o_{ki} (not a KRR*ki-tree node) is visited. $\{o_k, o_{ki}\}.\mathrm{score}$ is indicated as current best and o_k is the current best object. According to Preposition 4, any node$N_{ki} \in H$ is pruned if $\{o_k, N_{ki}\}.\mathrm{score} \leq \mathrm{current\_best}$. When H is empty, the current best object is $o_k.\mathrm{nn}_{ki}^1$. In the similar way, $o_k.\mathrm{nn}_{ki}^j$ (j > 1) can be identified.



Computing $lbkc_{t3}$

steps	Incrementally retrieve keyword-NN of t_3		S	kc.score
	"restaurant"	"hotel"		
1	$\{t_3, c_2\}$.score =0.7	$\{t_3, s_3\}$.score = 0.8	t_3, s_3, c_2	0.3
2		$\{t_3, s_2\}$.score = 0.28	$t_3, s_3, s_2, c2$	0.3
3	$\{t_3, c_3\}$.score =0.6		t_3, s_3, s_2, c_2, c_3	0.4
4	$\{t_3, c_4\}$.score =0.35		$t_3, s_3, s_2, c_2, c_3, c_4$	0.4

Fig. 4. An example of lbkc computation.

lbkc Computing Algorithm

Computing $\mathrm{lbkc_{ok}}$ is to incrementally redeem keyword-NNs of $\mathrm{o_k}$ in every non-main query keyword.

An example of lbkc computation. Is shown in figure 4 where Query keywords are "bar", "restaurant" and "hotel". The principal query keyword is "bar". Suppose we are computing $\mathrm{lbkc_{t3}}$. The first keyword-NN of $\mathrm{t_3}$ in "restaurant" and "hotel" are $\mathrm{c_2}$ and $\mathrm{s_3}$ respectively. A set S is used to keep $\mathrm{t_3}$, $\mathrm{s_3}$, $\mathrm{c_2}$. Let kc be the keyword cover in S which has the highest score (the idea of Apriori algorithm can be used, see section 5). After step 1,$\mathrm{kc.score} = 0.3$. In step 2, "hotel" is selected and the second keyword-NN of $\mathrm{t_3}$ in "hotel" is retrieved, i.e., $\mathrm{s_2}$. Since $\{\mathrm{t_3}, \mathrm{s_2}\}.\mathrm{score} < \mathrm{kc.score}, \mathrm{s_2}$ can be pruned and more importantly all objects not accessed in "hotel" can be pruned according to Preposition 5. In step 3, the second keyword-NN of $\mathrm{t_3}$ in "restaurant" is retrieved, i.e., $\mathrm{c_3}$. Since $\{\mathrm{t_3}, \mathrm{c_3}\}.\mathrm{score} > \mathrm{kc.score}, \mathrm{c_3}$ is inserted into S. As a result, kc is updated to 0:4. Then, the third keyword-NN of $\mathrm{t_3}$ in "restaurant" is retrieved, i.e., $\mathrm{c_4}$. Since $\{\mathrm{t_3}, \mathrm{c_4}\}.\mathrm{score} < \mathrm{kc.score}, \mathrm{c_4}$ and all objects not assessed yet in "restaurant" can be pruned according to Preposition 5. To this point, the current kc is $\mathrm{lbkc_{t3}}$.

Algorithm: $Local\_Best\_Keyword\_Cover(o_{k,}T)$
Input: A set of query keywords T, a principal object o_k
Output: $lbkc_{ok}$

```
1 foreach  $non - principal\ query\ keyword\ ki \in T$ do
2     $S$← retrieve $o_k.nn_{ki}^1$;
3     $k_i.score \leftarrow \{o_k, o_k.nn_{ki}^1\}.score$;
4     $k_i.n = 1$;

5 $kc$ ← the keyword cover in $S$;

6 while $T \neq \emptyset$ do
7     Find

$k_i \in T/\{k\}, k_i.score = \max_{k_j \in T/\{k\}}(k_j.score)$;
8     $k_i.n \leftarrow k_i.n + 1$;
9     $S$ ←$S$ ∪ retrieve $o_k.nn_{ki}^{ki.n}$;
10    $k_i.score \leftarrow \{o_k, o_k.nn_{ki}^{ki.n}\}.score$;
11    $temp\_kc$ ←the keyword cover in $S$;
12    if $temp_{kc}.score > kc.score$ then
13        $kc \leftarrow temp_{kc}$;
14        foreach $k_i \in T/\{k\}$ do
15            if $k_i.score \le kc.score$ then
16                remove $k_i$ from $T$

17 return $kc$;
```

It shows the pseudo-code of $\mathrm{lbkc_{ok}}$ computing algorithm. For each non-principal query keyword $\mathrm{k_i}$, the first keyword-NN of $\mathrm{o_k}$ is retrieved and $\mathrm{k_i.score} = \{\mathrm{o_k}, \mathrm{o_k.nn_{ki}^1}\}.\mathrm{score}$. They are kept in S and the best keyword cover kc in S is identified using Generate_Candidate function in Algorithm 3. The objects in different keywords are combined. Each time the most promising combination are selected to further do further combination until the best keyword cover is identified. When the second keyword-NN of $\mathrm{o_k}$ in $\mathrm{k_i}$ is retrieved, $\mathrm{k_i.score}$ is updated to $\{\mathrm{o_k}, \mathrm{o_k.nn_{ki}^2}\}.\mathrm{score}$, and so on. Each time one non-main query keyword is selected to search next keyword-NN in it. Note that we always select keyword k_i (∈ T) / {k} where k_i.score = max_(k_j ∈ T/{k}) (k_j.score) to minimize the number of keyword-NNs retrieved (line 7). After the next keyword-NN of $\mathrm{o_k}$ in this keyword is redeemed, it is inserted into S and kc is renovated. If $\mathrm{k_i.score}, \mathrm{kc.score}$, all objects in ki can be pruned by deleting $\mathrm{k_i}$ from T according to Preposition 5. When T is empty, kc is returned to lbkcok according to Preposition 6.

Keyword-NNE Algorithm

In keyword-NNE algorithm, the principal objects are processed in blocks instead of individually. Let k be the principal query keyword. The principal objects are indexed using KRR*k-tree.

Definition 6 (Corresponding Node): N_k is a node of KRR*k-tree at the hierarchical level i. Given a non-principal query keyword k_i, the corresponding nodes of N_k are nodes in KRR*ki-tree at the hierarchical level i.The root of a KRR*-tree is at hierarchical level 1, its child nodes are at hierarchical level 2, and so on. For example, if N_k is a node at hierarchical level 4 in KRR*k-tree, the corresponding nodes of N_k in keyword ki are these nodes at hierarchical level 4 in KRR*ki-tree. From the matching nodes, the keyword-NNs of N_k are redeemed incrementally for calculating $lbkc_{Nk}$.

Proposition 7: If a principal object ok is an object under a principal node N_k in KRR*k-tree

$$lbkc_{Nk}.score \geq lbkc_{ok}.score.$$

Proof: Suppose $lbkc_{Nk} = \{N_k, N_{k1}, \dots, N_{kn}\}$ and$lbkc_{ok} = \{o_k, o_{k1}, \dots, o_{kn}\}$. For each non-principal query keyword k_i, o_{ki} is under a corresponding node of N_k in keyword k_i are these nodes at hierarchical level 4, say N'_{ki}. Note that N'_{ki}can be in $lbkc_{Nk}$ or not. By definition,

$$lbkc_{Nk}.score \geq \{N'_k, N'_{k1}, \dots, N'_{kn}\}.score.$$

According to Proposition 2

$$\{N'_k, N'_{k1}, \dots, N'_{kn}\}.score \geq lbkc_{ok}.score.$$

So, we have

$$lbkc_{Nk}.score \geq lbkc_{ok}.score.$$

The Proposition is proved. The pseudo-code of keyword-NNE algorithm is presented in Algorithm 5. Keyword-NNE algorithm starts by selecting a principal query keyword k ∈ T (line 2). Then, the root node of KRR*k-tree is visited by keeping its child nodes in a heap H. For each node $\mathrm{N_k}$ in H, $\mathrm{lbkc_{Nk}.score}$ is computed (line 5). In H, the one with the maximum score, denoted as H.head, is processed. If H.head is a node of KRR*k-tree (line 8-14), it is replaced in H by its child nodes. For each child node $\mathrm{N_k}$, we compute $\mathrm{lbkc_{Nk}.score}$. Correspondingly, H.head is updated. If H.head is a principle object $\mathrm{o_k}$ rather than a node in KRR*k-tree (line 15-21), $\mathrm{lbkc_{ok}}$ is computed. If $\mathrm{lbkc_{ok}.score}$ is greater than the score of the current best solution bkc (bkc.score = 0 initially), bkc is updated to be $\mathrm{lbkc_{ok}}$. For any $\mathrm{N_k} \in \mathrm{H}, \mathrm{N_k}$ is trimmed if $\mathrm{lbkc_{Nk}.score} \leq$ bkc.score since $\mathrm{lbkc_{ok}.score} \leq \mathrm{lbkc_{ok}}$ for each $\mathrm{o_k}$ under $\mathrm{N_k}$ in KRR*k-tree according to Preposition 7. Once H is hollow, bkc is returned back to BKC query.

```
Input: A set of query keywords $T$, a candidate $can$, the current best solution $bkc$.
Output: A set of new candidates.
$New\_Cans \leftarrow \emptyset$;
$COM$ ←combining child nodes of $can$ to generate keyword covers;
foreach $com \in COM$ do
    if $com.score > bkc.score$ then
        $New_{Cans} \leftarrow com$;
return $New\_Cans$;
```

Algorithm: $Generate\_Candidate(T, can, bkc)$

WEIGHTED AVERAGE OF KEYWORD RATINGS

Till now, the minimum keyword rating of objects in O is used in $O.score$. However, it is not surprising that user prefers weighted average of keyword ratings of objects in O to measure $O.score$.

$$O.score = \propto \times \left(1 - \frac{diam(O)}{\max\_dist}\right) + (1 - \propto) \times \frac{W\_Average(O)}{\max\_rating} \qquad (8)$$

Where W_Average(O) is defined as [1]:

$$\mathrm{W\_Average(O)} = \frac{\sum_{\mathrm{o_{ki}} \in \mathrm{O}} \mathrm{w_{ki}} * \mathrm{o_{ki}.rating}}{|\mathrm{O}|} \qquad (9)$$

```
Input: A set of query keyword $T$, a spatial database $D$
Output: Best Keyword Cover
1  $bkc.score$ ←0;
2  $k$ ←select the principle query keyword from T;
3  $H$ ←child nodes of the root in KRR*k-tree;
4  foreach $N_k \in H$ do
5      Compute $lbkcNk.score$;
6  $H.head \leftarrow N_k \in H$ with $\max_{N_k \in H} lbkc_{Nk}.score$;
7  while $H \neq \emptyset$ do
8      while $H.head$ is a node in $KRR*k\text{-}tree$ do
9          N ←child nodes of $H.head$;
10         foreach $N_k$ in N do
11             Compute $lbkc_{Nk}.score$;
12             Insert $N_k$ into $H$;
13         Remove $H.head$ from $H$;
14         $H.head \leftarrow N_k \in H$ with $\max_{Nk \in H} lbkc_{Nk}.score$;
15 $o_k$ ←$H.head$;
16 Compute $lbkc_{Nk}.score$;
17 if $bkc.score < lbkc_{ok}.score$ then;
18     $bkc \leftarrow lbkc_{ok}$ ;
19 foreach $N_k$ in H do;
```

```
if $lbkc_{Nk}$.score ≤ bkc.score then;
    Remove $N_k$ from $H$;
return bkc;
```

where $\mathrm{w_{ki}}$i is the weight associated with the query $\mathrm{k_i}$ and $\sum_{\mathrm{k_i \in T}} \mathrm{w_{ki}} = 1$. For example, a user may give higher weight to "restaurant" but lower weight to "hotel" in BKC query. Given the score function in Equation (8), the baseline algorithm and keyword-NNE algorithm can be used to process BKC query. The necessity is to maintain the property in Proposition 1 and in Proposition 2. However, the property in Proposition 1 is not valid given the score function defined in Equation (9). To maintain this property, if a combination does not cover a query keyword $\mathrm{k_i}$, this is modified by inserting a virtual object associated with $\mathrm{k_i}$.

The $\mathrm{W\_Average(O)}$ is redefined to $\mathrm{W\_Average^{*}(O)}$.

$$\mathrm{W_{Average}}^{*}(\mathrm{O}) = \frac{\mathrm{E+F}}{|\mathrm{T}|} \qquad (10)$$

$$\mathrm{E} = \textstyle\sum_{\mathrm{o_{ki} \in O}} \mathrm{w_{ki}} * \mathrm{o_{ki}.rating.}$$

$$\mathrm{F} = \textstyle\sum_{\mathrm{k_j \in T/O.T}} \mathrm{w_{kj}} * \mathrm{O_{kj}.maxrating.}$$

where $\mathrm{T/O.T}$ is the set of query keywords not covered by O, $\mathrm{O_{kj}.maxrating}$ is the maximum rating of objects in $\mathrm{O_{kj}}$.

For example, suppose the query keywords are "hotel", "bar" and "restaurant". For a combination O ={t_1, c_1},$\mathrm{W\_Average(O)} = \mathrm{w_t} * \mathrm{t_1.rating} + \mathrm{w_c} * \mathrm{c_1.rating}$ while $\mathrm{W\_Average^{*}(O)} = \mathrm{w_t} * \mathrm{t_1.rating} + \mathrm{w_c} * \mathrm{c_1.rating}\ \mathrm{w_s}\ \mathrm{max\_rating}$ where $\mathrm{w_t}$, $\mathrm{w_c}$ and $\mathrm{w_s}$ are the weights assigned to "restaurant", "bar" and "hotel" respectively, and $\mathrm{max\_ratings}$ is the highest keyword rating of objects in "bar".

Given $\mathrm{O.score}$ with $\mathrm{W\_Average^{*}(O)}$, it is easy to prove that the property in Proposition 1 is valid. Note that the purpose of $\mathrm{W\_Average^{*}(O)}$ is to apply the pruning techniques in the baseline algorithm and keyword-NNE algorithm. It does not affect the correctness of the algorithms. In addition, the property in Proposition 2 is still valid no matter whether $\mathrm{W\_Average^{*}(O)}$ or $\mathrm{W\_Average(O)}$ is used in $\mathrm{O.score}$.

The keywords are combined to give candidate keyword covers which are stored in heap H. The candidate kc ∈ H with the maximum score is computed by retrieving the child nodes of kc. Then, the child nodes are combined to generate more candidates which replaces kc in heap H. This process stops when a keyword cover consisting of only objects is obtained. This keyword cover is the current best solution bkc. Any candidate kc ∈ H is pruned if $\mathrm{kc.score} < \mathrm{bkc.score}$. Once H is empty, the latest bkc is returned to BKC query.

APPLICATIONS

Online Yellow Pages- It allows users to specify an address and a set of keywords, and return businesses whose description contains these keywords, ordered by their distance to the specified address location.

Real estate websites- It allows users to search for properties with specific keywords in their description and rank them according to their distance from a specified location. We call such queries spatial keyword queries.

Subject search- If the user is not knowing about what to search for and in which category to search then subject search makes user query problem solve by providing respective search results.

CONCLUSION

Compared to the most relevant mCK query, BKC query provides an additional dimension to support more sensible decision making. The introduced baseline algorithm is inspired by the methods for processing mCK query. The baseline algorithm generates a large number of candidate keyword covers which leads to dramatic performance drop when more query keywords are given. The proposed keyword-NNE algorithm applies a different processing strategy, i.e., searching local best solution for each object in a certain query keyword. As a result, the quantity of candidate keyword covers produced is altogether decreased. The examination uncovers that the quantity of hopeful.

ACKNOWLEDGEMENT

We would like to express our deepest gratitude and sincere thanks to all those people who provided us the opportunity to complete the presentation. We are very grateful to our Principal Dr. R.I.K. Moorthy whose constant encouragement & motivation inspired us to do our best.

We would also like to thank our project guide Prof.Amol Kharat whose suggestions and encouragement helped us coordinate this presentation. We express our sincere thanks to the H.O.D of IT Dept. Mrs. Sharvari Govilkar and our project coordinator Mrs. Gayatri Hegde and Mrs. Sushopti Gawade whose encouragement made us do our presentation. Furthermore we would like to acknowledge the crucial role of the staff members of who gave us the permission to use all required equipment and the necessary materials, to complete our task.

REFERENCES

[1]Xin Li,Jiaheng Lu, Xiaofang Zhou “BEST KEYWORD COVER SEARCH” Knowledge and Data Engineering, IEEE Transactions on (Volume:27 ,Issue:1) MAY.2014

[2] Xin Cao et al. “Collective spatial keyword querying”.In:ACM SIGMOD 2015

[3] G. Cong, C. Jensen, and D. Wu. “Efficient retrieval of the top-k most relevant spatial web objects”. In: Proc. VLDB Endow. 2.1 (2009), pp. 337–348. 2012

[4] D. Papadias, N. Mamoulis, and Y. Theodoridis. “Processing and optimization of multiway spatial joins using R-trees”. In: PODS (2015), pp. 44–55.

[5] S. B. Roy and K. Chakrabarti. “Location-Aware Type Ahead Search on Spatial Databases: Semantics and Efficiency”.In: SIGMOD (2011).

[6] Dongxiang Zhang, Beng Chin Ooi, and Anthony K. H. Tung. “Locating mapped resources in web 2.0”. In: ICDE (2010).

[7] Dongxiang Zhang et al. “Keyword Search in Spatial Databases: Towards Searching by Document”. In: ICDE. 2009, pp. 688–699

[8] T.Brinkhoff, H. Kriegel, and B. Seeger. “Efficient processing of spatial joins using R-trees”. In: SIGMOD (2012), pp. 237–246.

[9] Ian De Felipe, Vagelis Hristidis, and Naphtali Rishe. “Keyword Search on Spatial Databases”. In: ICDE. 2013, pp. 656–665.

[10] Ramaswamy Hariharan et al. “Processing Spatial-Keyword (SK) Queries in Geographic Information Retrieval (GIR) Systems”. In: Proceedings of the 19th International Conference on Scientific and Statistical Database Management. 2007, pp. 16–23.

DESIGNING NEURAL NETWORK FOR IMAGE CATEGORIZATION

Parth Rajput ,Tejas Rahate ,Rahul Bhosale ,Kiran Nambiar
Department Of Information Technology
Pillai College Of Engineering ,New Panvel -410 206
University Of Mumbai
Academic Year 2017-2018

Abstract — This project explores the scope of Neural Networks in the field of Image Categorization. A Neural Network is a computing system that is biologically-inspired from a biological neuron, for the purpose of enabling a computer to learn from observational data. It was designed with the intent that it will learn to perform various tasks without providing a task specific code. In other words it can be said that a single neural network can be used in various applications with minimal changes. Neural networks provide the best solutions to many problems in image recognition, speech recognition, and natural language processing. This project demonstrates the use of a neural network to categorize images of Traffic Signs. It aims at implementing a neural network model and implementing it in an Android Application for real time detection of Traffic Signs. The key to the network, then, is figuring out effective transformations of the data to get good decisions. Due to the lack of Appropriate image datasets, a custom image dataset was created using images found online. These images were used to train and test our models to ensure maximum accuracy and low error rate. This project produces results that vary in accuracy based on the objects that need to be classified in the image and the type of image being used.

I. Introduction

The current projects on image categorization mainly make use of deep convolutional neural networks with many layers which make them difficult to deploy in embedded systems and other low resource areas, this project aimed at exploring the scope of using convolutional neural network model and retraining necessary layers. This project aimed at creating an optimized neural network model using Tensorflow. This project is aimed at using the optimized model for the purpose of real time traffic signs detection with an android application.

II. Objectives

The objective of this work is as follows:

A. To study and design a neural network using a robust and proven framework to categorize images based on various criteria (Road Signs Classification).

B. To implement a neural network to function in real time and predict accurate class labels for oncoming Road Signs while driving.

C. To train the network in a supervised way so that it is able to tag unclassified images.

III. Biological Neuron

As mentioned in the abstract a Neural Network, better known as Artificial Neural Network (ANN) is a computing system based on the biological neuron which is responsible for helping us compute from observed data. A neuron is a highly interconnected processing system. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly neurons working in accordance to solve specific problems. ANNs, like people, learn by example. A neural network is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of neural networks as well.

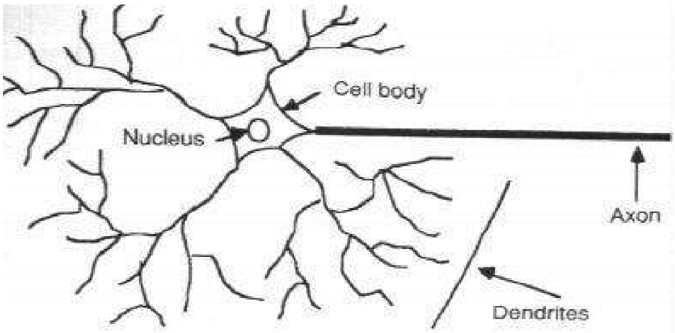


Fig. 1.1. Structure of a Neuron[1]

[11]The Figure 1.1 shows the human brain, a typical neuron collects signals from others through a host of fine structures called dendrites. The neuron sends out spikes of electrical activity through a long, thin strand known as an axon, which splits into thousands of branches. At the end of each branch, a structure called a synapse converts the activity from the axon into electrical effects that inhibit or excite activity from the axon into electrical effects that inhibit or excite activity in the connected neurons. When a neuron receives excitatory input that is sufficiently large compared with its inhibitory input, it sends a spike of electrical activity down its axon. Learning occurs by changing the effectiveness of the synapses so that the influence of one neuron on another changes.

IV. Literature Review

We have surveyed a variety of papers on image recognition. These papers discuss various convolutional networks and various related concepts. Our project is inspired by these concepts and makes use of convolutional networks for the purpose of image categorization.

Multi-GPU convolutional network(Krizhevsky et al., 2012)[3]: In this project a large, deep convolutional neural network was trained to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, the model achieved top-1 and top-5 error rates of 37.5% and 17.0% which was considerably better than the previous state-of-the-art image classification model. The neural network for this model contains 60 million parameters and 650,000 neurons, and it consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, non-saturating neurons were used along with a very efficient GPU implementation of the convolution operation.

COTS HPC[4] unsupervised convolutional network (Coates et al., 2013): In this project, deep learning algorithms have been scaled up which lead to an increase in performance of benchmark tasks and it also enabled discovery of complex high-level features. Recent efforts to train extremely large networks (with over 1 billion parameters) have relied on cloud-like computing infrastructure and thousands of CPU cores. In this paper, we present technical details and results from our own system based on Commodity Off-The-Shelf High Performance Computing (COTS HPC) technology; a cluster of GPU servers with Infiniband interconnects and MPI. This system is able to train 1 billion parameter networks on just 3 machines in a couple of days, and we show that it can scale to networks with over 11 billion parameters using just 16 machines. As this infrastructure is much more easily marshaled by others, the approach enables much wider-spread research with extremely large neural networks.

GoogLeNet[5] (Szegedy et al., 2014a): In this project, a deep convolutional neural network architecture codenamed Inception was proposed, which was responsible for setting the new state of the art for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14). The highlight of this architecture is the improved utilization of the computing resources inside the network. This was achieved by a cautiously crafted design that enables for increasing the depth and dimension of the network whereas keeping the machine budget constant. To optimize the standard, the architectural decisions were based on the Hebbian principle and the intuition of multi-scale processing. One particular version used in their submission for ILSVRC14 is called GoogLeNet, a 22 layers deep network, the quality of which is assessed in the context of classification and detection.

TensorFlow: Large-scale machine learning on heterogeneous systems(Martín Abadi, Ashish Agarwal, et al.,2015.) TensorFlow is a flexible data flow based programming model, as well as single machine and distributed implementations of this programming model. The system is borne from real-world experience in conducting research and deploying more than one hundred machine learning projects throughout a wide range of Google products and services. A TensorFlow computation is described by a directed graph, which is composed of a set of nodes. The graph represents a data flow computation, with extensions for allowing some kinds of nodes to maintain and update persistent state and for branching and looping control structures within the graph in a manner similar to Naiad

TABLE 2.1 Summary of literature survey

	Paper	Advantages and Disadvantages
1.	Multi-GPU convolutional network (Krizhevsky et al., 2012)	GPU clusters were used to accelerate computing and first time top 5 error rates were lower than existing technologies Advantage: 15.3 error rate, GPU introduced Disadvantage: significant performance degrade if network changed
2.	COTS HPC unsupervised convolutional network (Coates et al., 2013)	COTS can classify human faces, body(person) and animals requires GPU clusters for multiple days. Advantages: GPU acceleration efficiently implemented Disadvantage: Classification is limited to certain type of images.
3.	GoogLeNet (Szegedy et al., 2014a)	GoogleNet has 22 layers, can be re-trained to classify wide range of objects Advantages: Optimized computing, scalable, efficient GPU utilization Disadvantage: Designed for competition hence retraining is limited to types of classes in IMAGENET dataset.
4.	TensorFlow: Large-scale machine learning on heterogeneous systems(Martín Abadi, Ashish Agarwal, et al.,2015.)	Advantages: structure of the computation graphs and behavior of machine learning models is understood using a visualization tool called TensorBoard. Disadvantages: Applications are practically limited to deep learning since it lacks solid understanding of mathematical concepts and a steep learning curve where other methods are more profound .

V. CATNET-Categorization Network

Figure 3.1 shows the approach being used to implement the system. This architecture gives a brief explanation of the working of the proposed system. As seen in the flowchart, an image that is to be categorized will first be pre-processed. This would include various image enhancement operations and removal of background noise. As in most systems, the partitions(segments) of the image that is significant for categorization, are segmented for greater accuracy.

Based on these partitions, features will be extracted from them. This can be understood with the example of the image of a "No Parking" road sign. The features that may be extracted from such an image would be the shape, color, text, etc. As various images are bound to have some number of similar features, they are processed to determine which features define a particular class.

In the next stage, the image is finally classified. the image is labelled in various possible classes along with their probabilities. The confidence for the classification is calculated and compared with the results of classification of the image with other standard models. Based on this comparison a deviation is calculated to determine the performance of the model. All the results of image classification and deviation is displayed to the user.

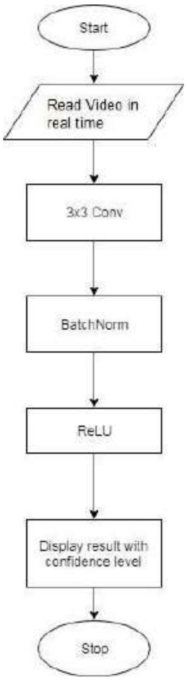


Fig. 3.1. Flowchart of CatNet

A. Existing System Architecture

Definition of classes: Main responsibility of the component is to represent the types of road signs (e.g.regulatory signs, warning signs, guidance signs, etc) coming from various information sources and datasets.

Feature selection: The module collects data and tries to extract the features of the road signs. Usually, this is achieved through machine learning techniques, which are able to extract features based on the shape of the road sign, its colour,etc.

Sampling of training data: The module from the knowledge known from the features which were extracted , tries to identify various other road signs. It is the most crucial component in identification process as good training of data results in better results.

Estimation of statistics: It is responsible for the output from the previous step is analysed and the a estimation is made for which sign it is depending upon the support and confidence level percentage. For better accuracy it is required that ample amount of training is done on testing dataset.

Classification: It depends on the various estimation given, the given road sign image is classified into the various classes of road sign.

Verification of results: The final step in the process of identification of the image. Based on the class the system has classified and the support and confidence level shown the user verifies the given result.

B. System Architecture and Implementation

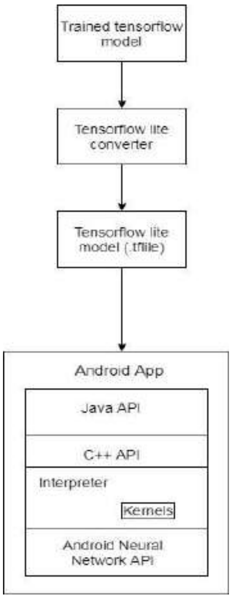

Trained tensorflow model
Tensorflow lite converter
Tensorflow lite model (.tflite)
Android App
Java API
C++ API
Interpreter
Kernels
Android Neural Network API

Fig. 3.2 Architecture of Tensorflow Lite

The previous system requires the user to supply a pre-processed dataset. It required image processing[4] to be performed and supply the processed images as input to the neural network. This saved the need for extra convolution layers but increased the overhead of processing the images before using them. In order to save this overhead we allow input of colored images. In order to achieve better domain results, we used the existing system architecture as a template to build upon and make suitable changes to achieve the necessary functions and add new advantages to the system while removing the previous disadvantages.

First our android application detects the object of interest i.e. the traffic signs within the images that are sampled in real time by the application using the camera of the mobile device. Then the image that is visible on the viewport of the android application is recognised by the neural net working behind the scene. The app displays the result of the image that is identified with the confidence level.The confidence level is displayed between 0 - 1, 0 being the lowest and 1 being the highest. The recognition of the image is based on the tensorflow for poets codelab which is set of tutorials that helps implement the Mobilenet and Tensorflow Lite class of models . The Mobilenets are optimized to be small and efficient, at the cost of some accuracy. While the model was initially designed for Mobilenet it was later also translated for Tensorflow Lite due to its higher optimization.

C. Sample Dataset Used

To classify images we needed a training set of images.We created a dataset of images for various traffic signs. At least 100 copyright free images for each traffic sign were downloaded using extension. These images were then manually examined for their usability and they were put under folders named after the respective sign. Initially small datasets for two traffic signs were created. Gradually both the number of images and traffic signs were increased and various trends were observed in the accuracy of the model based on the images and these changes in the dataset.

Table 3.4. Sample Dataset Used for Experiment

Dataset	**Items**	**Type**
Street Sign India	500	Information

D. Hardware and Software Specifications

The experiment setup is carried out on a computer system which has the different hardware and software specifications as given in Table 3.2 and Table 3.3 respectively.

Table 3.5. Hardware details

Processor	4 GHz Intel 4790
SSD	120 GB
RAM	16 GB
GPU	GTX 970 4GB

Table 3.6. Software details

Operating System	Ubuntu 14/ Windows 10
Programming Language	Python 2.7, Java.
Platform	Android
IDE	Android Studio
Frameworks	Tensorflow

E. Evaluation Metrics

The quality of the model is indicated by the score that is displayed for classification of each image. This score is an indication of confidence. It is evaluated with the help of confidence score.

Confidence : Confidence indicates the number of times the if/then statements have been found to be true.

$$\mathbf{conf}(X \Rightarrow Y) = \mathbf{supp}(X \cup Y)/\mathbf{supp}(X).$$

Eq. 3.2. Confidence

X=>Y - X and Y are features and X implies Y

VI. Applications

A. Traffic and Road sign recognition

This system first searches signs within images captured by the sensor on the vehicle, and then identifies road signs helping the driver of the vehicle to properly drive the vehicle. It identifies between the different types of road signs based on shape of the sign (circle,triangle, square) and the shapes inside the recognised border . Besides it can be used to recognize traffic. The model can be used to recognize vehicles a single person or a crowd on the road.

B. Military Application

With the ability to detect the objects in real-time, it can be used in robots or drone to identify the targets or to identify the campsites for delivery of essential supplies to the soldiers

.

C. Remote sensing

Remote sensing relies to the interaction of electromagnetic radiation with matter. In remote sensing, fuzzy neural networks have been used for a variety of applications such as military reconnaissance, flood estimation, crop prediction, mineral detection, and oil exploration . Active systems such as SAR can infiltrate clouds that block the view of passive systems, like the multispectral and panchromatic sensors.

D. Google Net

The image processing method are also implemented using the mobileNet model which classifies images based on classes determined by training dataset (CIFAR-10) and extracts relevant features to determine what class does the image belong to. This model minimises human error which may occur while specifying classification features .

E. Automated taxi routing

Similar system is applied in vehicle stabilization while automated driving . Google and Tesla have adapted the technologies to center the car according to the lane borders tracked using the image processing models and using neural networks to predict the required position of the car and setting the car variables accordingly .Such an applications are also applied to valet parkings and automated taxi routing .

F. Real-Time Object Detection

YOLO[7] is a real-time object detection system. On a Titan X, it processes images at 40-90 FPS. It uses a single neural network to the full image. This network divides the image into regions and predicts bounding boxes and probabilities for each region. These bounding boxes are weighted by the predicted probabilities. Hence similar method is used to implement image processing over videos .

VII. Summary

In this report, the study of different approaches for Categorization of Images is presented. These approaches make use of the concepts of Neural networks, of which models such as convolutional networks have been discussed. Various other approaches have also been mentioned. The comparative study of various techniques mentioned above is presented in this report. For this project, Image Categorization is being implemented with the help of a convolution network with multiple layers. The network uses multiple layers which are found to be useful for image classification techniques viz. Convolution Layer, Rectified Neuron Layer and Max Pooling Layers and Softmax. The performance measures are the accuracy and confidence score, which are described in this report. The different standard datasets or variable inputs are defined that may be used in experiment for this domain systems. One of the datasets identified for this project is the Government dataset for Road-signs. The applications of this domain is identified and presented.

In our project, the dataset has been generated once, saved and used every time to train the neural network. The network can be retrained using different dataset of street signs. The model is lightweight and can be used on embedded systems and mobile applications, it requires less computing resources compared to other popular convolution models and the accuracy of this model will be compared with that of other popular models. The end product is an android application, displaying all the necessary details such as classified labels, support confidence and comparative study of other models which can be used for a variety of purposes.

Acknowledgment

We sincerely express our gratitude to our Mentor Prof. Varunakshi Bhojane for guiding us through our project. Her counsel and advice have made it easy for us to achieve our goals in a more informative and productive manner. We further thank the HOD Dr. Sharvari Govilkar and Dr.Madhumita Chatterjee, the faculty for allowing us to pursue this project and also in helping us to guide and decide between the plethora of options the college had to offer .We would also like to thank Dr. Sandeep Joshi providing the required resources and guidance for the project , without his support this project would not have been possible and we are very grateful for his encouragement.

References

[1] Sasikumar Gurumurthy, Balakrushna K. Tripathy, M. Priya, “Study of Image Recognition Using Cellular Associated Artificial Neural Networks”, (https://www.researchgate.net/figure/50864266_fig1_Fig-1-A-Simple-Neuron)

[2] Quoc Le Tomas Mikolov, “Distributed Representations of Sentences and Documents”

[3] Alex Krizhevsky, et al. “ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems 25” ,NIPS 2012- A large, deep convolutional neural network to classify the 1.3 million high-resolution images in the LSVRC-2010 ImageNet training set into the 1000 different classes was trained.

[4] Adam Coates, Brody Huval, Tao Wang, David J. Andrew, “Deep learning with COTS HPC systems.”

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, “Going Deeper with Convolutions.”

[6] Chen, Huizhong, et al. "Robust Text Detection in Natural Images with Edge-Enhanced Maximally Stable Extremal Regions." Image Processing (ICIP), 2011 18th IEEE International Conference on. IEEE, 2011.

[7] Karl Moritz Hermann, et al. “Teaching Machines to Read and Comprehend “- Machine reading systems can be tested on their ability to answer questions posed on the contents of documents that they have seen, but until now large scale training and test datasets have been missing for this type of evaluation..

[8] Jeff Donahue, Yangqing Jia, et al. “DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition”, researchgate.net- It evaluates whether features extracted from the activation of a deep convolutional network trained in a fully supervised fashion on a large, fixed set of object recognition tasks can be repurposed to novel generic tasks

[9] YOLO: Real-Time Object Detection (https://pjreddie.com/darknet/yolo/).

[10] Brandon Rohrer, How Convolutional Neural Networks Work,(https://www.kdnuggets.com/2016/08/brohrer-convolutional-neural-networks-explanation.html?utm_source=mybridge&utm_medium=email&utm_campaign=real_more)

[11] Geoffrey E. Hinton, “How Neural Networks Learn from Experience”

DETECTION AND REMOVAL OF A FAKE PRODUCT REVIEWS

Prof. Shubhangi Chavan,Abhishek Kadupatil, Akshay Gawade, Shrihas Devalekar, Rohan Warkhade
srathod@mes.ac.in, akadupatil@student.mes.ac.in , akshay781@student.mes.ac.in,
shrihas167@student.mes.ac.in , rwarkhade@student.mes.ac.in
Department of Information Technology
PILLAI COLLEGE OF ENGINEERING
New Panvel

ABSTRACT:

Trust plays an important role in any commerce transactions and especially in E commerce transactions. Lack of trust is the main reason why people fear to buy products on e-commerce websites. So,it is the biggest task to remove fear from the minds of buyers and improve the site's reputation. The only way to solve this problem is by building trust in their minds.

Trustworthiness is a very critical element and requires evaluation of a wide variety of information types,parameters and uncertainties. In this project, we ensure that users give genuine reviews on the products. Trust Reputation System(TRS) algorithm provides a most Trustful reputation score for a specific product so as to support customers to take right decision while interacting with e-commerce website. TRS relies on an appropriate algorithms which improves the selection,generation and classification of textual feedbacks. This algorithm studies the customer's attitude towards selection of prefabricated reviews. After doing this process, the reputation algorithm generates better trust degree of the user, trust degree of the review and a better global reputation score of the product. This ensures that the product sold online get the perfect rating according to its capabilities and in turn helps customers to make a right choice about the product.

I. INTRODUCTION

“What other people thoughts are and their thinking” has always been an important source of information for most of us during the decision-making process. Long before awareness of the World Wide Web (www) became widespread, many of us requested our friends to recommend a mixer or to explain who they were thinking to vote for elections, requested reference letters regarding job applicants from friends, or consulted Consumer Reports to decide what mixer to buy. With the rapid expansion of e-commerce, many products are sold on the Web, and many people are also buying products online. In order to enhance customer satisfaction, requirements and online shopping experience, it has become a common practice for online merchants to enable their customers to suggest opinions on the products that they have purchased. With more and more common users becoming comfortable with the Web, a growing number of people are writing reviews and posting them which are becoming beneficial for others. As a result, the number of reviews that a product receives grows rapidly. Some popular products can get hundreds of reviews at large merchant sites. And our application will give you the promising reviews by filtering them from other sites. And then you can decide what you want to buy or not.

II. LITERATURE SURVEY

In 2017, Prof.Manleen Kaur Kohli and Prof.Shaheen Jamil Khan proposed **“Fake Product Review Monitoring and Removal for Genuine Online Product Reviews Using Opinion Mining” [1]**. They mentioned that With the rapid expansion of e-commerce, many products are sold on the Web, and many people are also buying products online. In order to enhance customer satisfaction, requirements and online shopping experience, it has become a common practice for online merchants to enable their customers to suggest opinions on the products that they have purchased. With more and more common users becoming comfortable with the Web, a growing number of people are writing reviews and posting them which are becoming beneficial for others. As a result, the number of reviews that a product receives grows rapidly.

In 2017 **,** Prof. Sandeep Yadav and Prof. Mohit Sinha proposed **“Online Fake Product Review & Monitoring” [2]**. In that system, they have observed that Some fake reviews could be written with malicious intentions to distort the reputation of businesses on E-commerce websites. For example, employees could post fake positive reviews to praise offerings of their organizations, and fake negative entries to chastise those of rival businesses.This Paper describes the way to detect users malicious intent by Decision tree Algorithm called as “Decision Tree-Classification(ID3 Algorithm)” The basic strategies used by Id3 algorithm to first choose the splitting attribute which is having highest information gain.That is called root node.Flowcharts are used in designing and documenting complex processes or programs. The two most common types of boxes in a flowchart are: a processing step, usually called activity, and denoted as a rectangular box a decision, usually denoted as a diamond. This algorithm does the work of tracking the IP and Validate the review.It is upon the admin to track those fake reviews.

In 2017 **,** Prof. Pan Liu,Prof. Zhenning(Jimmy) Xu proposed an **“Identifying indicator of fake reviews based on smmer’s behaviour features” [3]**. They have noticed that online data is crucial is E-marketing business,they play a huge role in defining the quality of the product,However Spam or Fake reviews are affecting the reliability of data analysis and decision making,enterprise. To detect spam reviews, the paper presents a set of opinion spam detection identification indicators based on behavior features of the spammer.This paper deals with behavioural pattern of spammers by techniques called “Star User”,”Deviation Rate”,”Bias Rate”,”Review Similarity rate”,”Review RelevancyRate”,”Content-Length”,and”Burst length”.It also consists of various algorithms like “Conversation Times”,”LCS”,” Calculate Similarly” and “Word Segmentation” which deals with various opinion spamming process to deal with fake comments.

In 2015**,** Prof. Ankita Thakkar and Prof.Deepali Vora proposed for **“Building trustworthiness of user-feedbacks on products in online shopping environment” [4]**. In that they mentioned,e-com is a very effective strategy for online shopping,but there is no direct communication with a vendor so there is a biggest challenge is to remove fear from mind of a buyer so only solution is to build trust in their mind for this we use a technology of Sentiment Analysis to know the sentiment behind the given semantic feedback and check its concordance with rating given by the same user. If concordance is satisfactory user has to like or dislike five prefabricated feedbacks. Using both the concordance and like-dislike results, Trust Degree of the user is calculated.the Trust Degree is above certain threshold then the feedback of that user is submitted to our website else the user is blocked. This is how we improve website quality by our implementation.

In 2015, Prof. Reena Mahe and Prof. Rahul Jadhav proposed a system on **“Trustworthiness in e-commerce context using trs algorithm” [5]**. In that they have mentioned, trust is an important factor for carrying E commerce activities.Lack of trust may ruin the transactions or even lower the reputation of the product or the retailer.This paper introduces the use of Trust Review System (TRS) Algorithm through which malicious interventions of users is detected whose intention is to falsify the Reputation score of a product positively or negatively.The main idea of TRS is to detect the user's intention towards the prefabricated feedbacks provided it.The user first login to the E-Commerce site from which he wants to purchase a product, then after seeing it the user provides its feedback about the product. The feedback is then stored in the ordinary database. Those feedbacks can be fabricated in order to summarize numerous users’ feedbacks which are stored in from which the database.The generated feedbacks can also be stored in another knowledge base. So as much as we add feedbacks in the ordinary database,TRS Algorithm is applied in which user has to like or dislike prefabricated reviews which calculates the trust score of user through which fake comments can be detected.

In 2015, Prof. Daya Mevada and Prof. Viraj Daxini proposed **"An opinion spam analyzer for product reviews using supervised machine learning method" [6]**. In that they have mentioned, While making any purchase online consumer usually checks opinions of others about the product. Manufacturer can gain insight into its products strength and weaknesses based on the reviews of the customers. So, user reviews play a crucial role in Web, since many decisions are made based on them. However usefulness of this reviewing systems motivates some people to enter their fake review to promote some products or defame some others. Opinion spam analyzer, which classifies reviews into non-spam reviews and spam reviews, and provide quality data to review mining these opinion spam should get detected and eliminated in order to prevent misleading potential customers. In this thesis, opinion spam detection approaches have been proposed and examined over a sample dataset. The proposed system will use supervised machine learning method to classify opinions into spam or non-spam. we proposed approach for performing spam detection in opinion reviews by merging methods from machine learning and text mining in one classification approach.

.

III.PROPOSED SYSTEM ARCHITECTURE

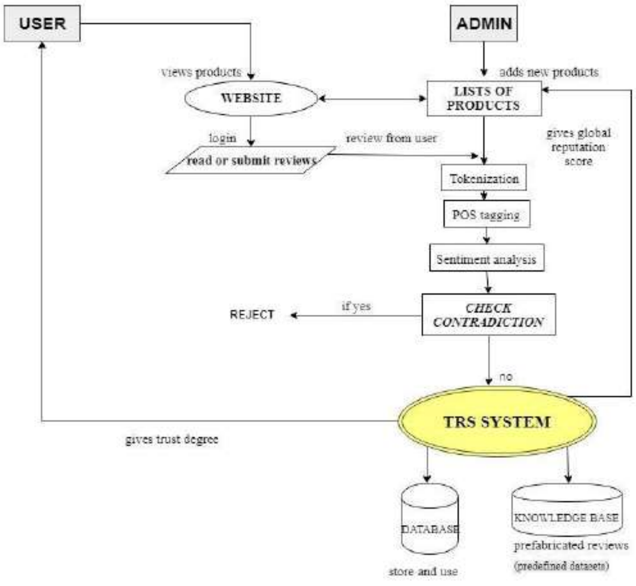


Fig 2. Detection of Fake Product Reviews System

In the proposed system, we are creating a comparative website of electronic appliances where the admin will be adding the product details and their reviews collected from five different e-commerce websites and user visiting our website can view those reviews. If the user wants to submit reviews on our website , his/her review genuinity will be analysed by the **TRS System.** The working of **Trust Reputation System** algorithm in our project is given as follows :

Proposed TRS System aims at creating trust and propagating it in online communities while giving actionable results .Those results such as global trust degree, trust scores and Sentiment Analysis help users to make a decision about purchasing or not a particular product. Proposed design will use both ratings and especially semantic feedbacks in order to calculate trust weight and to classify comments and users. Sentiment Analysis will be performed on Feedbacks. Feedbacks of the users will only be submitted to portal if the trustworthiness of the user is above certain threshold value.

The user starts by giving an appreciation (rating) and a textual feedback about a specific product. When he clicks on submit in order to validate the given information, we are going to redirect the user to another interface showing this message for example: "please give us your opinion about the following feedbacks before validating the information you gave below:"

In this interface we will find chosen feedbacks from the database from different types. Those feedbacks can be fabricated in order to summarize numerous users‟ feedbacks stored in the database. The generated feedbacks can be stored in another knowledge base. So as much as we add feedbacks in the ordinary database, we will fill the knowledge data base with prefabricated feedbacks using text mining algorithms and tools. However, some users can give already summarized feedbacks that can directly be included in the knowledge database.

Actually, before sending the user's feedback and appreciation about the product to the Trust Reputation System, we have to verify the concordance between them in order to avoid and eliminate contradiction or malicious programs attacking our system. In the redirected interface, we will display several feedbacks from different types. However, the user can specify the number of feedbacks to be liked or disliked. Of course, we can also specify the minimum and the maximum number of feedbacks to be displayed by the user.

We are trying through this redirection to detect and analyze the user intention behind his intervention on our comparative website. Hence, we examine and evaluate his intention using other prefabricated feedbacks with different types. Of course, we have already the trustworthiness of each feedback. Consequently, we use our Reputation algorithm studied in order to generate the user trust degree which plays the role of a coefficient and then rectify his appreciation according to his trust degree and generates the score of the feedback.

Indeed, each feedback has trustworthiness in a threshold [-5,5]. The closest is the trustworthiness to 5, the most trustworthy the feedback is. The closest is the trustworthiness to -5, the very untrustworthy is the feedback. If the feedback is trustworthy its score would be included in [0,5] else it would be included in [-5,0].

After that, we have to generate the global trust reputation score of the product using the user's appreciation (rating) and his trust degree. In fact, a possible example for such a rating method might be school marks and coefficients. By doing this, TRS system tries to build confidence in the user and improves the trustworthiness of our comparative website. The user can finally trust on our reviews before buying.

IV. TECHNIQUES

1. Data Preprocessing

Data preprocessor component aims to collecting and cleaning the data before subsequent analysis. Spidering programs will be used to collect the source data from websites as HTML pages. When data are gathered from websites it contains HTML tags and other non-textual data. opinion Spam classification textual task usually requires text data in the form of comments. The data preprocessor module removes such non textual contents and gives the data in structured text format for using in opinion spam classification. The preprocessor removes noisy characters from the input documents.

2. Tokenization

Tokenization process splits the text into very simple units such as numbers,punctuation and words.These tokens are often loosely referred to as terms or words, but it is sometimes important to make a type/token distinction. A token is an instance of a sequence of characters in some particular that are grouped together as a useful semantic unit for processing.

3. POS tagging

Not all the words in source text data are useful for sentiment analysis. The part-of-speech (POS) tagging assigns each token a tag which may be adjective, verb, or adverb. Liu et al. founds that nouns and noun phrases in the sentences are the features of the products and adjectives and adverbs are used to express opinions through opinion words. Part-of- speech (POS) tagging is useful for identifying adjectives and adverbs in the sentences which identify the opinion words, and nouns which identify the features of the products.

4. Trust Reputation System(TRS)

Reputation algorithm used in this TRS is using sentiment feedbacks analysis in order to generate a trustful reputation score for the User. Actually, we have 3 types of feedbacks:

Positive Feedbacks: represent opinions that expressing a positive point of view about the product. Those ameliorative opinions contain a positive content concerning the product.Then, the adjective positive is referring to the nature of the content of the feedbacks not its trustworthiness. However, each feedback whatever is its type can have either a positive trustworthiness or a negative trustworthiness. Either positive trustworthiness or negative one, it is gradual: it has degrees as float in a threshold of [-5.5].

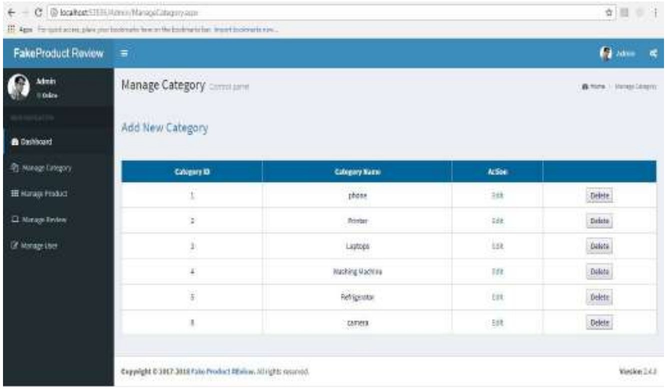
Negative Feedbacks: represent opinions talking negatively about the product. Logically, the users giving such opinions are not satisfied of the commented product. This feedback could be telling the truth or a part from the truth or could be far from the truth. That's why; each feedback has its trustworthiness represented by a float number between -5 and 5.

Contradictious Feedbacks: represent feedbacks with a contradictious content for example a feedback where the user is not talking about the specified product but another one or he/she is affirming that the camera of a mobile phone is great and later in the same opinion is saying that the camera is very bad. In fact, we have to start by detecting the contradictious feedbacks. Then we are in need of a sentiment analysis algorithm and tool that can detect the contradiction in a specific content related to a product. We can personalize the analysis according to the product. For instance, if the user says that "the swimming pool of the hotel which doesn't afford one is not clean", the algorithm must be able to detect this great contradiction. We can give to the algorithm for each product as an input the property of the algorithm; if there is no similarity we can consider it as a contradiction.

V. RESULTS

GUI of Admin Side

5.1 Manage Categories



Manage Category Control panel

Add New Category

Category ID	Category Name	Action	
1	phone	Edit	Delete
2	Printer	Edit	Delete
3	Laptops	Edit	Delete
4	Washing Machine	Edit	Delete
5	Refrigerator	Edit	Delete
6	camera	Edit	Delete

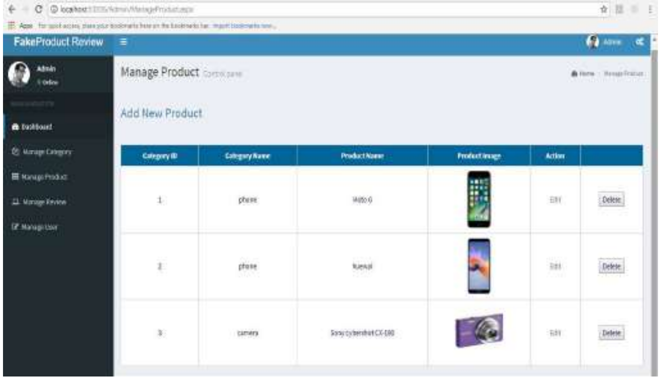
Copyright © 2017-2018 Fake Product REview. All rights reserved.

Version 2.4.0

figure 5.1 manage categories

The GUI shown in figure 5.1 shows the admin added new categories into the system and can edit or delete the existing categories from the system.

5.2 Manage Products



Manage Product Control panel

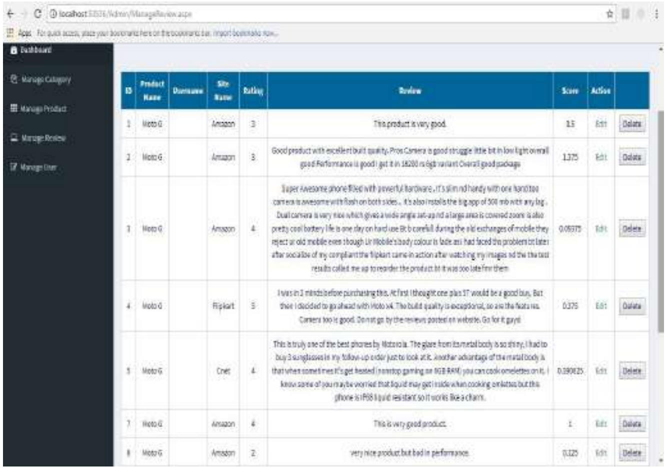
Add New Product

Category ID	Category Name	Product Name	Product Image	Action	
1	phone	Moto G		Edit	Delete
2	phone	Huawei		Edit	Delete
3	camera	Sony cybershot CX-180		Edit	Delete

figure 5.2 manage products

The GUI shown in figure 5.2, the admin adds new products by selecting category , adds description, price and image of the product.

5.3 Manage Reviews

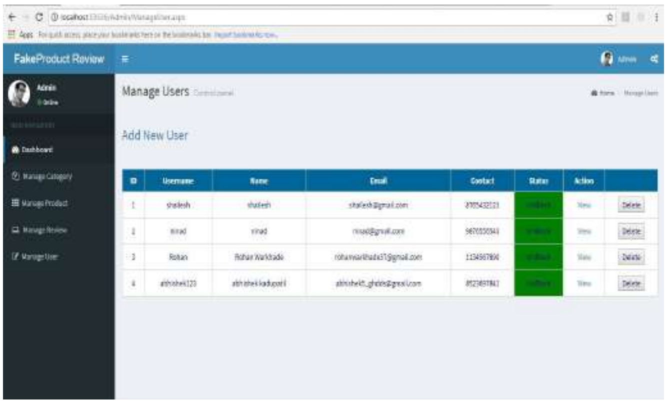


ID	Product Name	Username	Site Name	Rating	Review	Score	Action	
1	Moto G		Amazon	3	This product is very good.	1.5	Edit	Delete
2	Moto G		Amazon	3	Good product with excellent built quality. Pros Camera is good struggle little bit in low light overall good Performance is good I got it in 18200 rs 6gb variant Overall good package	1.375	Edit	Delete
3	Moto G		Amazon	4	Super Awesome phone filled with powerful hardware , it's slim nd handy with one hand too camera is awesome with flash on both sides .. it's also installs the big app of 500 mb with any lag . Dual camera is very nice which gives a wide angle set-up nd a large area is covered zoom is also pretty cool battery life is one day on hard use Bt b carefull during the old exchanges of mobile they reject ur old mobile even though Ur Mobile's body colour is fade as i had faced the problem bt later after socialise of my compliant the flipkart came in action after watching my images nd the the last results called me up to reorder the product bt it was too late frm them	0.05375	Edit	Delete
4	Moto G		Flipkart	5	I was in 2 minds before purchasing this. At first I thought one plus 5T would be a good buy, But then I decided to go ahead with Moto x4. The build quality is exceptional, so are the features. Camera too is good. Do not go by the reviews posted on website. Go for it guys!	0.375	Edit	Delete
5	Moto G		Cnet	4	This is truly one of the best phones by Motorola. The glare from its metal body is so shiny, I had to buy 3 sunglasses in my follow-up order just to look at it. Another advantage of the metal body is that when sometimes it's get heated (nonstop gaming on 8GB RAM) you can cook omelettes on it. I know some of you maybe worried that liquid may get inside when cooking omlettes but this phone is IP68 liquid resistant so it works like a charm.	0.190625	Edit	Delete
7	Moto G		Amazon	4	This is very good product.	1	Edit	Delete
8	Moto G		Amazon	2	very nice product but bad in performance.	0.125	Edit	Delete

figure 5.3 manage reviews

The GUI shown in figure 5.3 shows the admin can add reviews to the products by taking reviews from five ecommerce websites for that same product and generate the score through sentiment analysis of the review. TRS system generates the score of the review by performing tokenization, POS tagging on the review, applying sentiment analysis on the review with the help of SentiwordNET. The admin can also view the reviews added/submitted by the registered user on our website.

5.4 Manage Users



ID	Username	Name	Email	Contact	Status	Action	
1	shailesh	shailesh	shailesh@gmail.com	[illegible]	[illegible]	View	Delete
2	vinad	vinad	vinad@gmail.com	[illegible]	[illegible]	View	Delete
3	Rohan	Rohan Warkhade	rohanwarkhade31@gmail.com	[illegible]	[illegible]	View	Delete
4	abhishek123	abhishek kadupatil	abhishek_ghlds@gmail.com	[illegible]	[illegible]	View	Delete

figure 5.4 manage users

The GUI shown in figure 5.4 shows the admin can view all the user registered with the website, view their details and status of block/unblock. The admin can also view the blocked user by the TRS system and can also unblock the block user on user's request cross-checking his/her trust score.

GUI of User Side

5.5 Home Page

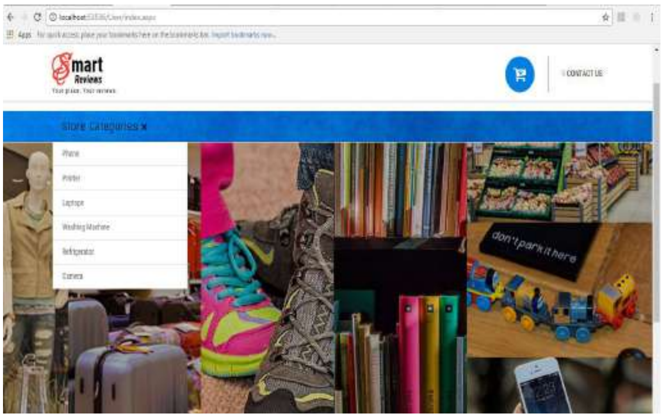


figure 8.5 home page

5.6 View Products

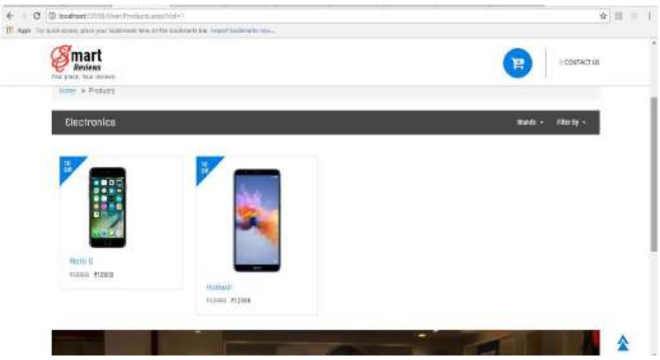


figure 5.6 view products

The GUI shown in figure 5.6 shows the user can view products by selecting a category, read specifications and view price and mainly the reviews.

5.7 View Reviews

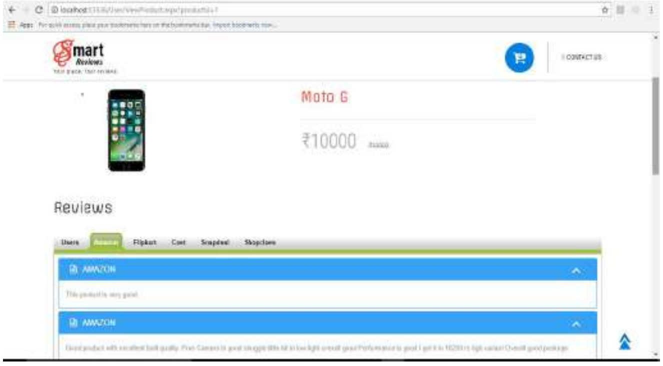


figure 5.7 view reviews

The GUI shown in figure 5.7 shows the user views all the reviews of a particular product selected to view and can submit rating and review after login.

5.8 Submit Review

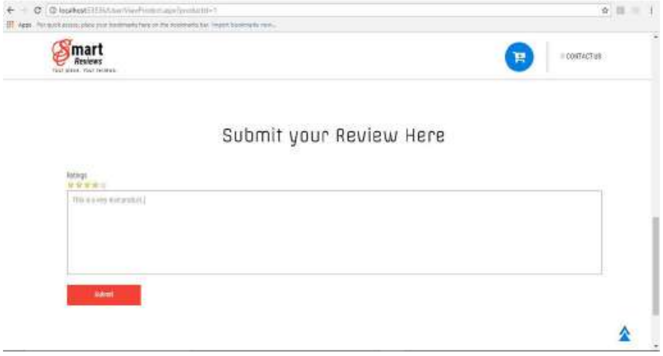


figure 8.8 submit review

The GUI shown in figure 5.8 shows the registered user is able to submit the review for the product after login to the website. If the user is not registered, then he/she needs to register first. Then only further activity can be done.

5.9 Login Page

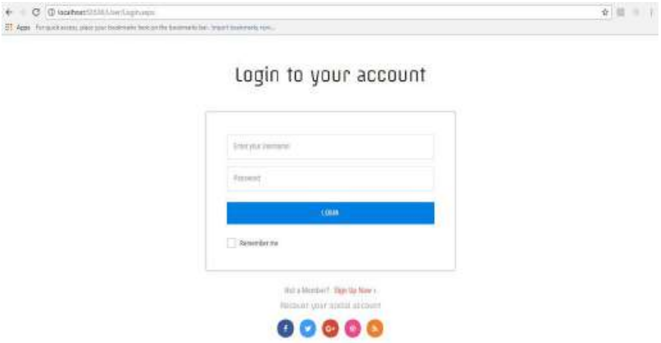


figure 5.9 login page

The GUI shown in figure 5.9 shows the user logs into the website with valid username and password. After successful validation of credentials of the user, user can do the activity of submitting the review.

5.10 Registration Page

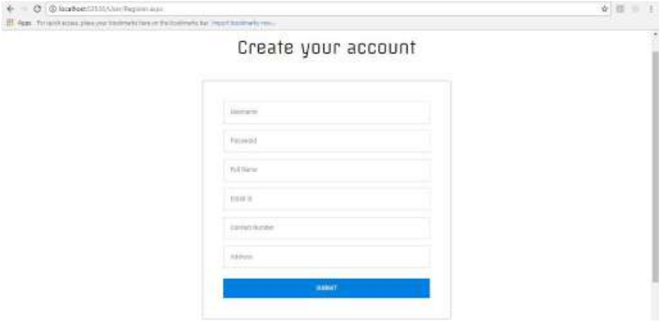


figure 5.10 registration page

The GUI shown in figure 5.10 shows the user needs to enter valid details for the registration process .

5.11 Like-Dislike Page

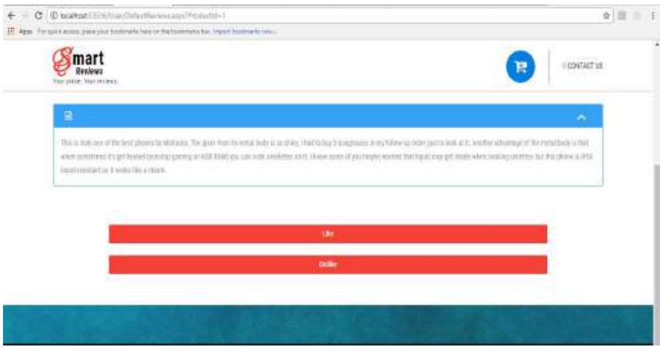
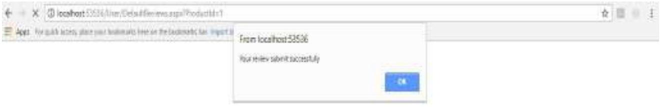


figure 5.11 like-dislike page

The GUI shown in figure 5.11 shows the newly registered user is redirected by the system before submitting the review. For the product. Depending on the like-dislike action by the user on this page, the new user is assigned a trust score. This trust score is important for the system to judge the genuinity of the user and its review. After successful activity, the user will be able to submit the review for the product. If system finds contradiction, the user may get blocked.

5.12 Submission of review


From localhost:53536

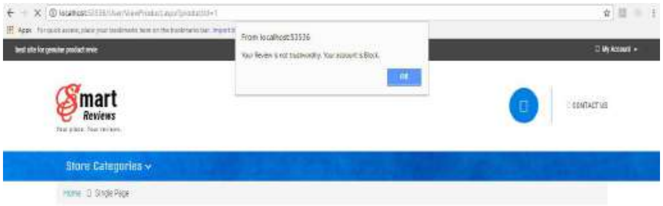
Your review submit successfully

figure 5.12 submission of reviews

After the successful process done in figure 5.11, the GUI shown in figure 5.12 shows the user’s review gets submitted and this message is shown to the user.

5.13 Blocking of Untrusted User

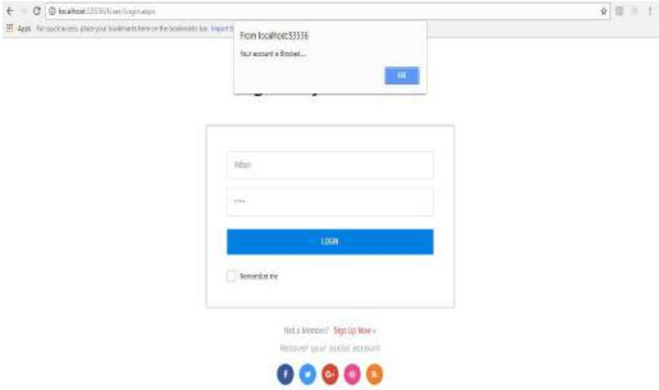
The GUI shown in figure 5.13 shows the message to the user that he/she is blocked. This is because the TRS system determines that the trust score of the user falls below the threshold and the account of that user needs to be blocked as the reviews submitted the user may be fake or prefabricated.


From localhost:53536

Your Review is not trustworthy. Your account is Block.

figure 5.13 blocking of untrusted user

5.14 Blocked User trying to Login to the System



From localhost:53336

Your account is Blocked...

figure 5.14 blocked user trying to login to the system

The GUI shown in figure 5.14 shows that when the blocked user tries to login into the system with the username and password, he/she will not be able to login and the system will show the message that your account is blocked. That user will be able to login again only when the admin unblocks the user.

VI. CONCLUSION AND FUTURE SCOPE

The study of different domain techniques is presented. The different techniques such as Tokenization, Part of speech tagging, Sentiment analysis, Contradiction Verification. Trust Reputation System aims on creating trust score of the product and proposing it online to help user in transactions. These rating and trust scores helps the user in making a decision whether to buy a product or services online by the specific vendor. TRS ensures that false review and grading does not be displayed, it does it by discarding it.Some of the algorithms have been used in Trust analysis gives good results.This approach is based on an intelligent layer that semantically analyses the user's feedback to determine its sentiment about the product and its trustworthiness.

In the future, we would focus on genuinity of the product and improving the user's experience on the system. We could focus on increasing the accuracy of our trust reputation system for a larger dataset and add-on new technologies based on machine learning to reduce the workload of the admin. Also, our TRS system was unable to handle the negation in the sentences, so further we would do research on how to tackle the issue and will try to implement it.

VII. REFERENCES

[1] Prof. Pan Liu,Prof. Zhenning(Jimmy) Xu,Prof. Jun Ai, ”Identifying Indicators of Fake Reviews Based on Spammer's Behavior Features”,October 2017 in IEEE International Conference on Software Quality, Reliability and Security (Companion Volume),2017 IEEE DOI 10.1109/QRS-C.2017.72.

[2] Prof. Sandeep Yadav ,Prof. Mohit Sinha , Prof. Dilip Kesari , Prof. Dilip Kesari, ” Online Fake Product Review & Monitoring”,March 2017 in International Journal of Research In Science & Engineering, Special Issue 7-ICEMTE, e-ISSN: 2394-8299,p-ISSN: 2394-8280.

[3] Prof.Manleen Kaur Kohli, Prof.Shaheen Jamil Khan, Prof.Tanvi Mirashi, Prof.Suraj Gupta ,“Fake Product Review Monitoring and Removal for Genuine Online Product Reviews Using Opinion Mining”,January 2017 in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 1, ISSN: 2277 128X.

[4] Prof. Ankita Thakkar, Prof.Deepali Vora ,”Building Trustworthiness of User-Feedbacks on Products in Online Shopping Environment”, April 2015 in International Journal on Recent and Innovation Trends in Computing and Communication,Volume: 3 Issue: 4,ISSN: 2321-8169,1770 – 1772.

[5] Prof. Reena Mahe, Prof. Rahul Jadhav, Prof.Pratik Gaikwad, Prof.Rahul Gadekar, Prof.Kiran Bhise,“Trustworthiness in E-Commerce Context using TRS Algorithm”,October 2015 in International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 10,ISSN (Online) 2278-1021, ISSN (Print) 2319 5940.

[6] Prof. Daya L. Mevada, Prof.Viraj Daxini, ”An Opinion Spam Analyzer For Product Reviews Using Supervised Machine Learning Method”, November 2015 in Journal of Information, Knowledge and Research in Computer Engineering,VOLUME – 03, ISSUE – 02,ISSN: 0975 – 6760.

[7] Kishori K. Pawar, Pukhraj P Shrishrimal, R. R. Deshmukh,”Twitter Sentiment Analysis: A Review”,april 2015,International Journal of Scientific & Engineering Research, Volume 6, Issue 4,ISSN 2229-5518.

[8] Hasnae RAHIMI, Hanan EL BAKKALI,” New Reputation Algorithm for Evaluating Trustworthiness in E-Commerce Context”, 2015

[9] www.technopedia.com/uml-diagram

[10] www.slideshare.net/

[11]www.cs.sjsu.edu/~pearce/oom/requirements/use-cases.htm

[12] https://en.wikipedia.org/wiki/Reputation_system

[13]https://www.sciencedirect.com/science/article/pii/S0167923605000849

ENHANCED DATA ENCRYPTION USING HYBRID CRYPTOGRAPHY

Prof. K. S. Charumathi, Kunal Kadam, Sunita Pawar, Paresh Jha, Kanchan Chaube

Department of Information Technology,
Pillai's College of Engineering New Panvel, Maharashtra, India.

kscharumathi@mes.ac.in
kadamka15it@student.mes.ac.in
pawarss15it@student.mes.ac.in
jhapd15it@student.mes.ac.in
chaubeka15e@student.mes.ac.in

Abstract—Cryptanalyst are expert in how to break the encryption techniques. We need to safe our programs and documents from cryptanalyst. Security of information means protecting data from unauthorized access in cloud environment. There are many techniques to achieve the security of information from unauthorized access. There are two cryptographic techniques used for data encryption which are Symmetric and Asymmetric techniques. There are some advantages and disadvantages of these block cipher algorithms. Rijndael had a potentially lower security margin and better performance than being arguably simpler than many encryption techniques. Symmetric key encryption algorithms are computationally fast compared to asymmetric encryption algorithms (like RSA). However, since the same secret key is used for symmetric encryption and decryption, we have the difficult problem of securely distributing that secret key. Conversely, asymmetric key infrastructure in PKI does not rely on distribution of any private key. However, the common asymmetric algorithms are too slow to be used for bulk encryption with current computation capability. While SHA is better collision resistant to various block cipher algorithms. Thus for better security performance, we propose a system which would incorporate the advantages of these algorithms namely SHA, AES-RIJNDAEL and RSA which will be a hybrid approach of encrypting data. SHA is adopted in this mechanism to verify the integrity of the message. Three major security principles such as authentication, confidentiality and integrity are achieved together using this scheme.

***Keywords*— Hybrid cryptography, RSA, DES, Rijndael, AES, Security, Data Security, Encryption, SHA512, Hashing.**

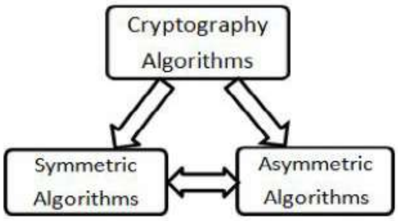
I. Introduction

With advancements in computer technology and the widespread reach of the world-wide-web i.e. the Internet, people lives are changing rapidly. The liberalization, internationalization and personalization features of Internet have been attempting to bring revolutionary reform to government agencies, enterprises and institutions, at the same time help to boost the work efficiency and market response to improve their competitiveness by the use of Internet. But how to make the information system confidential and make sure that it is not leaked, even if they are stolen it is difficult to be identified, if they are identified after all the difficulty, they are extremely difficult to be modified. To prevent confidential information from being accessed, modified, and fabricated, to keep that protected has become a hot research topic in the IT industry.

In today's times Cloud computing has a significant impact on the IT industry. With growing popularity more and more organizations are making use of cloud services. Although cloud services have a widespread acceptance but the fear pertaining to security and privacy of these services still continue to be an open challenge. With rapid technological advancements these services could be easily accessed through smart phones thus allowing users to share pictures, video, documents and other important data across various platforms on a real time basis. However, a security breach in there Security has always been a concern in the domain of information technology. With Cloud services handling critical data which can be accessed from anywhere through the internet makes security a prominent concern. The pervasive nature of Cloud and its disbursal of data across various geographical locations amount to high security risks. While talking of Cloud Security there are many aspects which one needs to consider such as, trusted authentication, appropriate authorization, data security and privacy. These are some of the basic security goals which are extremely essential for every cloud provider to incorporate. Since security has been seen as an attribute for information technology, data encryption has been one of its key measures in ensuring data security protection.

II. Hybrid Cryptography Description

The encryption technology (Cryptography) is the basic safety techniques used in current e-commerce and banking websites which are of extreme importance. Information encryption technology can not only meet the security requirements of confidentiality of information, but also avoid the leakage of the important information which are of high security especially in the security (defence) and hospital, banking sectors. Therefore, encryption technology is the base of authentication technology, as well as many other security technologies that are used today. Cryptography is an art of writing and reading the secret information. It uses mathematics in science to protect the information. It is a method of encrypting the original information into a form that is not easily interpreted by anyone. Original message can be revealed only after decrypting the encrypted message. Public and private keys are used for this purpose. Generally, the cryptographic systems can be classified into symmetric and asymmetric. In symmetric cryptography, same key is used for the encryption and decryption whereas in asymmetric cryptography separate keys are used for the encryption and decryption process. There are two types of cryptography algorithm that are given below:



- Symmetric key cryptography algorithm
- Asymmetric key cryptography algorithm

To increase the security level, this proposed scheme overcomes the limitation of “Basic encryption algorithm proposed till date. The proposed enhanced scheme includes RIJNDAEL, RSA and SHA. RIJNDAEL (Variant of AES) strengthens the security of data stored in cloud. Reason behind for selecting RIJNDAEL rather other encryption algorithms is that, the key used for encryption and decryption is suspected to meet-in-middle attack. RSA is used to solve the key distribution problem and in addition to this, SHA to verify the integrity of the data. Use of SHA algorithm in combination of cryptographic algorithm provides strength in security of data stored in cloud.

Due to small key size DES is insecure and has weaknesses. Triple DES which is an enhancement to DES, the original DES algorithm was applied thrice to increase the security. But it was found to be very slow. Blowfish algorithm runs faster than other symmetric algorithms. AES algorithm is the best encryption algorithm. The blowfish algorithm is fastest as compare to other algorithms but it has less security than the AES.

III. Purpose Of Hybrid Cryptography

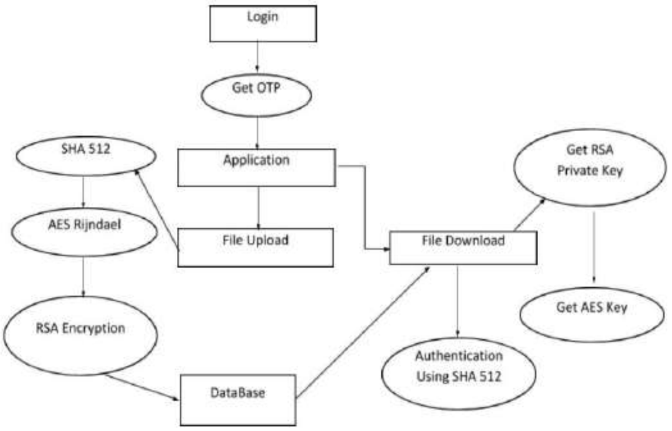
The confidentiality, integrity and availability of resources are three major issues in this cloud computing security. IT infrastructure developers are eager to deal with gradually increasing secure algorithms in cloud networks. Still the area of cloud is open for data security in cloud network and seeking for more reliable, secure and less complex model. The security in the field of cloud being more improved when the attribute based encryption implemented in cloud data. Where the encryption of data with the key and that key is encrypted with adopted attribute, the whole combination of ciphertext, key and attribute combine to become master key. Hybrid Encryption plays an important role in mitigating risk related to the many threats listed in this guide. If sensitive information stored on your computer is encrypted, it will take a secret key to decode it. If sensitive information in route to others is encrypted, only someone that knows the secret key can read what it says. When you encrypt sensitive information and it ends up logged by others in the course of communicating online, encryption keeps those without the secret key from knowing the contents of the message. Most of the Defensive Technology articles will cover practical ways to apply encryption to particular communications (like email) or particular applications (like web browsers).Encryption is absolutely essential to maintaining information security. Moreover, modern computers are powerful enough that we can aim to make encryption of our communications and data routine, not just reserving encryption for special occasions or particularly sensitive information.

IV. Existing System

In the current encryption systems, individual algorithms are used to secure data. Such as Linux systems use MD5 encryption algorithm while some others use maybe AES or DES algorithms to encrypt their passwords. But each of these mentioned algorithms have been cracked some or the other time, which means they are not invincible and can be broken by a skilled hand. Thus the security of the data (passwords in many cases) is highly and threateningly compromised. All these algorithms are very famous all around the globe and are used by many, some are even open source. This means that the algorithm’s flaws are well known to all and in some cases, even the source code is well known to many. This adds up to the security woes of these algorithms. Thus there needs to be a system which overcomes these drawbacks while upholding the positive aspects of these widely known algorithms.

V. The Proposed System

To increase the security level this proposed scheme overcomes the limitation of "Hybrid encryption algorithm proposed by Wuling Ren. The proposed enhanced scheme includes Rijndael, RSA and SHA-1. Rijndael (Variant of AES) strengthens the security of data stored. Reason behind for selecting Rijndael rather than other is that the key used for encryption and decryption is suspected to meet-in-middle attack. RSA is used to solve the key distribution problem and in addition to this, SHA-1 to verify the integrity of the message. Use of message digest SHA-1 algorithm in combination of cryptographic algorithm provides strength in security of data storage in cloud. Here we specify different modules of envision system.



VI.Conclusions

In this report we have stated how we are going to work on making our system more secure using hybrid cryptography. Having security aspect in mind we have discussed algorithms such as Rijndael algorithm, RSA algorithm, SHA-512 algorithm. We have discussed how these algorithms would be implemented in the proposed system. We have described main drawbacks that are present in the present system and how those issues can be resolved to an extend using hybrid cryptography which is our proposed system.

VII. Acknowledgement

Professor K.S.Charumathi has guided us throughout our project and we would like to offer our deepest appreciation for his efforts and patience. We would like to acknowledge and appreciate Head of Department. Sharvari Govilkar for her support throughout the project. Our final appreciation is directed towards our Principal R.K Moorti for always keeping his doors open for consultation or advice required.

VIII.References

[1] Thakur Jawahar, Kumar Nagesh. "DES, AES and Blowfish Symmetric Key Cryptography algorithm Simulation Based Performance Analysis", IJETAE, vol.1, Issue 2, DEC. 2011, pp.6-12.

[2] Jignesh R Patel, Rajesh S. Bansode Vikas Kaul, "Hybrid Security Algorithms for Data Transmission using AES-DES" International Journal of Applied Information Systems (IJAIS), Volume 2– No.2, February 2012.

[3] Jigar Chauhan , Neekhil Dedhia, Bhagyashri Kulkarni , International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 3, May 2013. Enhancing Data Security by using Hybrid Cryptographic Algorithm.

[4] Lalit Singh Dr. R.K. Bharti, "Comparative Performance Analysis of Cryptographic Algorithms" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013.

[5] Meenakshi Shankar and Akshaya.P , International Journal of Network Security & Its Applications (IJNSA) Vol.6, No.6, November 2014. Hybrid Cryptographic Technique Using RSA Algorithm and Scheduling Concepts.

[6] Jitendra Singh Laser, Viny Jain, "A Comparative Survey of various Cryptographic Techniques" International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 03 | Mar-2016.

[7] Md. Alam Hossain, Md. Biddut Hossain, Md. Shafin Uddin, Shariar Md. Imtiaz, "Performance Analysis of Different Cryptography Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering Volume 6, Issue 3, March 2016.

Product Suggestion to Customers using Sentiment Mining with the help of Scrutiny of Products

Ritesh Naik, *Student, PCE*,Sandesh Auti , *Student, PCE*,Aniket More, *Student, PCE*,Nirbhay Mhatre, *Student, PCE*, and Mrs.Deepti Lawand. Faculty, *PCE*

***Abstract*—With the Internet becoming more universal,e-commerce gradually on the rise,businesses have a stores entities & creating virtual store that is shopping websites on the internet.When consumers search for their desired product on multiple shopping websites,they need to filter and compare search results and compare different price on different shopping websites by themselves.Therefore,it always takes lots of time for the consumers,and even the search results do not accord with consumers demand.This tool proposed to resolve all the problems problems from consumer side and will also help for E-commerce shopping websites to understand choice of consumers. Admin will vary the price of products based on overall reviews about that particular product or else can add/remove products.A large amount of database will also be taken for consideration.Consumer can add reviews,comments,stars,etc.NLP(Natural Language Processing) use to read scrutiny and use of naive bayes classification to check statistics of reviews.**

I. Introduction

E-Commerce sites pervade the internet. A wide variety of products are sold online including electronic goods, apparel and household items. With mobile phones becoming a common medium of accessing the internet, m-commerce too is gaining rapid momentum. India is one of the fastest growing E-Commerce and E-Retailing markets with the market expected to grow to around USD 9 billion by 2016. With such a rapid growth in this industry, companies are using sophisticated algorithms to understand the buying patterns of their buyers in order to enrich the customer experience. There is cut throat competition among E-Commerce sites in the way they present their products, the promotions and discounts they offer and the shopping experience they provide to customers. These offerings are based on extensive market research and analytics conducted by experts within and outside these organizations. One of the key parameters that companies use to strategize is customer reviews and rating on the e-commerce sites. These reviews are not only used by the companies but also play a major role in consumers deciding whether to buy a product or not. Hence analyzing customer reviews help both shoppers as well as E-Commerce companies.

For individual customers, it could be a cumbersome process to read through each review of various products and make decisions. For instance, on the website of the India E Retailing company Flipkart, recently launched Redmi 1S phone has nearly 4800 ratings and 3900 reviews. There are numerous mobile phones with very similar features and in such circumstances customers rely on the reviews of others before making a decision. Hence E-Commerce sites provide as many details of reviews as possible on their websites.

While making their decision, customers look at the following aspects:

Number of star ratings

Positive and Negative tone of reviews

Various features of products (eg. Battery life, RAM, screen resolution with respect to mobile phones) discussed in reviews

Helpfulness factor of reviews

Authenticity of reviews

Number and age of reviews

1.1 MOTIVATION

The overall rating of the product given by e-commerce users does not provide other customers with a clear understanding of the user’s perspective on feature wise performance of the product. Also, business managers do not get an accurate insight of how these ratings affect the sales of the product. Although several methods have been proposed to use product reviews for business intelligence, limited work has been done on using customer satisfaction as a metric. Prior works had customer satisfaction ratings and recommendations independent of each other. Hence the focus is on building a framework for measuring customer satisfaction which can help the business administrators to take strategic decisions as well as provide the customers with valuable infographics which would help them in taking informed decisions regarding which product should be purchased and give them recommendation on the basis of the generated satisfaction rating. The aim is to extract patterns and develop them into insights based on aspect based sentiment analysis on product reviews aggregated from the e-commerce sites and present solutions by visualizations for the management and satisfaction based product recommendation for customers so as to enhance transactions on these sites.

1.2 PROBLEM STATEMENT

E-Commerce sites are gaining popularity across the world. People visit them not just to shop products but also to know the opinion of other buyers and users of products. Online customer reviews are helping consumers to decide which products to buy and also companies to understand the buying behavior of consumers. When consumers search for their desired product on multiple shopping websites,they need to filter and compare search results and compare different price on different shopping websites by themselves.Therefore,it always takes lots of time for the consumers, and even the search results do not accord with consumers demand.This tool proposed to resolve all the problems problems from consumer side and will also help for E-commerce shopping websites to understand choice of consumers. In our proposed system we extract the reviews of product features and the polarity of those features. We graphically present to the customer, the better of two products based on various criteria including the star ratings, date of review, the helpfulness score of the review and the polarity of reviews.

1.3 SCOPE

To mine the opinions of the people sentiment analysis is the best approach. The purpose of this mining review is to benefit the customers and encourage them to buy products online without facing any problem. In this comparison of two products is done as an extension it will be more than two with specifications and also with the product based system of products.

II.LITERATURE SURVEY

As we know that web engineering is a field in which web data is processed in different way.In past, lots of organization might have been done research about those technique.Literature review for Product recommendation system possess some of the big data analysis technique.We shall study about the previous researches done on those techniques.

In some studies where opinion mining used but they don't include security mechanism of user login and Collaborative Filtering for filters information by using the recommendations of other people[3].In order to achieve better results we have included both techniques i.e.Naive Bayes Classification and Collaborative Filtering along with that inclusion of security mechanism will help us to build flawless proposed systems, which seek to inherit vantages and eliminate disadvantages.

2.2 Literature Review

Venkata Rajeev P & Amrita Vishwa Vidyapeetham, "Recommending products to customers using opinion mining Of online product Reviews & features", International Conference on Circuit,Power & Computing Technologies, 2015.

In this paper, Venkata Rajeev P & Amrita Vishwa [1] have created a prototype web based system for recommending and comparing products sold online. They have used natural language processing to automatically read reviews and used Naive Bayes classification to determine the polarity of reviews. They have also extracted the reviews of product features and the polarity of those features. They graphically present to the customer, the better of two products based on various criteria including the star ratings, date of review, the helpfulness score of the review and the polarity of reviews.

Ming-Hsiung Ying, Yeh-Yen Hsu,"A Commodity Search System for Online Shopping Based on Ontology and Web Mining"IET The Institution of Engineering & Technologies, 2014.

In this paper,Ming-Hsiung Ying, Yeh-Yen Hsu[2] attempts to use semantic analysis, ontology, and web mining technique as a basic approach. This study proposes a novel commodity search system to track consumer demand, and that is, when the commodity price of any website is lower than the consumer price conditions, the system will proactively notify consumers.This study designed three different uses of the agent to aid in searching commodities.The commodities information crawl agent will download commodities saved in the database, so that consumers can search commodities on this study system.

Miss Lovenika Kushwaha, Prof. Sunil Damodar Rathod,"New Opinion Mining Technique for Online Product Reviews and Features", Multidisciplinary Journal of Research in Engineering and Technology, Volume 2, Issue 4, Pg.852-858.

In this paper, Miss Lovenika Kushwaha, Prof. Sunil Damodar Rathod[3] opinion mining is used to process the online product reviews, feature and recommend the best product among others. In this paper they have created a prototype web based system for recommending and comparing products which sold online on websites.They have also extracted the reviews of product features and the polarity of those features. This study results indicate that the novel product search system could assist consumers to search commodity, and provide historical price information of commodity for consumers to decide.

Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7(1):76–80, 2003.

At Amazon.com, they use recommendation algorithms to personalize the online store for each customer.The store radically changes based on customer interests, showing programming titles to a software engineer and baby toys to a new mother. The click-through and conversion rates two important measures of Web-based and email advertising effectiveness vastly exceed those of untargeted content such as banner advertisements and top-seller lists.

Recommendation algorithms provide an effective form of targeted marketing by creating a personalized shopping experience for each customer. For large retailers like Amazon.com, a good recommendation algorithm is scalable over very large customer bases and product catalogs, requires only subsecond processing time to generate online recommendations, is able to react immediately to changes in a user's data, and makes compelling recommendations for all users regardless of the number of purchases and ratings[4].

III.Advanced Product Recommendation System

3.1 Overview

Social networking and e-commerce sites provide the opportunity for people to interact with each other and publicly share their opinions about other people, places, products and events. A platform is provided to express opinions quantitatively through scores, star ratings or votes as well as qualitatively through text and videos. The internet is now filled with such opinions and will serve as a "gold mine" to companies trying to under their customers. When customers write reviews of products, most of them focus on specific aspect of the product. For example, "Screen Resolution is poor", "Battery drains too fast" or "Excellent audio quality" are some reviews commonly written for mobile phones. Hence it is not just important to get an overall idea of the review but also to understand what features customers are satisfied with and what features make customers unhappy[4].

This feature based extraction is of immense benefit to both customers and sellers who are looking for making improvements to the product as well as marketing strategies.

In our proposed system we extract the reviews of product features and the polarity of those features[3]. We graphically present to the customer, the better of two products based on various criteria including the star ratings, date of review, the helpfulness score of the review and the polarity of reviews.

3.1.1 Existing System Architecture

The process of mining the opinions or views of the users for the products or services they have used and detecting the orientation of the sentiment of the sentence is sentiment analysis. Sentiment analysis means to infer the opinion polarity of the review i.e. deciding whether the opinion expressed is positive or negative or neutral.

Applications of sentiment analysis are found in product reviews, stock market prediction, election results predictions and political debate analysis.

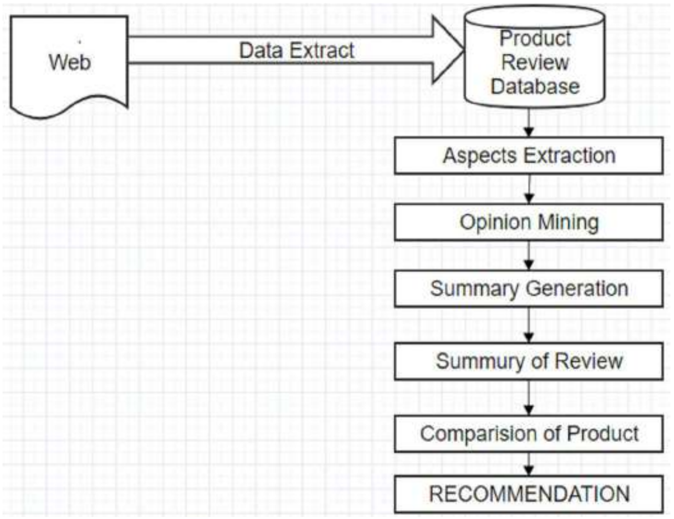


Fig. 3.1 Existing system architecture[3]

Key Components of Existing System architecture

Web: Web contains the information about product.This data is required for recommendation system.

Product Review Database: The data is stored at one location known as database.Product reviews are stored here.

Aspect Extraction: All those data related aspects are extracted for further extraction.

Opinion Mining: The system uses opinion mining methodology in order to achieve desired functionality. Opinion Mining for Comment Sentiment Analysis is a web application which gives review of the topic that is posted by the user.

Summary Generation: Summary of opinion mining is generated by using text compactor.

Summary of Review: Summary of reviews are calculated by considering the reviews given by total number of users.

Comparison of Product: Products are compared according to price and polarity of reviews.

Recommendation: Finally Product is review to user.

3.1.2 Proposed System Architecture

The previous sections discussed the strengths and weaknesses of existing system. In order to achieve better domain results, researchers combined both techniques to build Hybrid domain systems, which seek to inherit vantages and eliminate disadvantages.

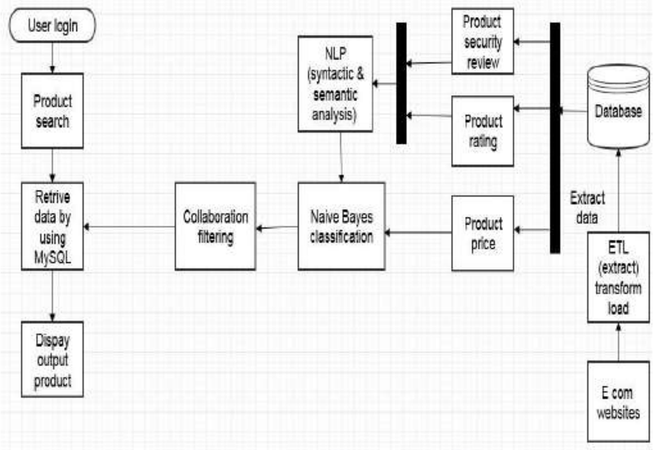


Fig. 3.2 Proposed system architecture

Components of proposed system architecture

Ecommerce website: Ecommerce websites are the main backbone of our proposed system.Two or more than two ecommerce website will be taken for this.For ex:-Amazon, Flipkart,ebay.

Database:The data of e commerce website will be stored in database by using ETL processing.This database contains product reviews, rating,price.

Product Scrutiny/Review: Product review is the feedback of customer.This reviews are in comments.

Product Rating: Product rating is the symbolic feedback given by customer.This rating is usually given in 5-star rating model.

Product Price: Cost of product is an important factor in Product suggestion system.

Naive Bayes Classification: Processed data is classified by using Naive Bayes Classification technique.

Collaborative Filtering: It filters information by using the recommendations of other people.This filtering is not present in the existing system.

User Login: The user has to login into the system and then can he make use of the system resources. The user need not login all the time; once he's logged in he is remembered until he logs out.The user data is validated by administrator.

Retrieve Data: Finally the output data will be retrieve to user.MySql language is used for retrieval of data

3.2 Requirements for Implementation

The implementation detail is given in this section.

3.2.1 Techniques

ETL(Extract, Transform & Load)

ETL is short for extract, transform, load, three database functions that are combined into one tool to pull data out of one database and place it into another database. Extract is the process of reading data from a database. In this stage, the data is collected, often from multiple and different types of sources. Transform is the process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data.Load is the process of writing the data into the target database.

Syntactic & Semantic

To converse with humans, a program must understand syntax (grammar), semantics (word meaning), morphology (tense), pragmatics (conversation). The number of rules to track can seem overwhelming and explains why earlier attempts initially led to disappointing results[1].

Syntactic Analysis (Parsing) − It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among the words. The sentence such as “The school goes to boy” is rejected by English syntactic analyzer.

Semantic Analysis − It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain. The semantic analyzer disregards sentence such as “hot ice-cream”.

Collaborative Filtering

Collaborative filtering, also referred to as social filtering, filters information by using the recommendations of other people. It is based on the idea that people who agreed in their evaluation of certain items in the past are likely to agree again in the future[2]. A person who wants to see a movie for example, might ask for recommendations from friends. The recommendations of some friends who have similar interests are trusted more than recommendations from others. This information is used in the decision on which movie to see.

Security Mechanism

As we are going to introduce user registration and login page,it is mandatory to provide security to user's private data.In context to this,we are going to provide encryption technique to the user's login password.Techniques use for database security:-

MD5 Encryption:

MD5 stands for ‘Message Digest Algorithm 5’.MD5 algorithm is used as a cryptographic

Hash function for a file fingerprint.Often used to encrypt password in database,MD5 can also generate a fingerprint file to ensure that file is same after a transfer for example.A MD5 hash is composed of 32 hexadecimal characters.

3.2.3 Sample Dataset Used

An experiment is conducted in order to identify the input/output behavior of the system. Identify inputs. Specify the sample inputs that would be used in the experiments. The sample dataset used in the experiment are identified and given in Table 3.1

Table 3.1 Sample Dataset Used for Experiment.

Products	Number of Reviews(out of 5 crore)	Number of Ratings
Books	8,898,041	22,507,155
Headphones	1,689,188	7,824,482
Shoes	1,697,533	4,607,047
RAM	1,097,592	3,749,004
Jewelry	278,677	5,748,920
Home & kitchen	551,682	4,253,926
Laptop	982,619	3,205,467
T-Shirt	296,337	3,268,695
Ear ring	194,439	3,447,249
Comfort sets	346,355	2,982,326
USB	167,597	2,252,771
Toys and games	231,780	1,324,753
Tools & Home improvement	134,476	1,926,047
Beauty	198,502	2,023,070
Apps for Android	752,937	2,638,172

IV.IMPLEMENTATION

4.1 Evaluation Parameters

- Deterministic:-Deterministic functions always return the same result any time they are called with a specific set of input values and given the same state of the database.
- Non Deterministic:-Nondeterministic functions may return different results each time they are called with a specific set of input values even if the database state that they access remains the same.
- Effectiveness:-It is the capacity of algorithm to handle the amount of data in number.Some algorithm may have small capacity to handle the data or some algorithm may have large capacity to handle the data.
- Speed:-It is the amount of time taken by algorithm to give result.
- Accuracy:-It is the ability of algorithm to obtain a good result from a large number of dataset.

4.2 Performance Evaluation

- Performance evaluation is done on the basis of certain parameters.Those parameters are given below on the basis of those parameters it is confirmed that Decision tree gives best result.

Table 4.2 Parameter Difference

Parameter	**Decision Tree**	**Naive Bayes**
Deterministic/ Non-Deterministic	Deterministic	Non-Deterministic
Effectiveness on	Large Data	Huge Data
Speed	Faster	Slower than Decision tree
Accuracy	High accuracy	For obtaining good results it requires a very large number of records

- **Accuracy(in percentage):-**We firstly find out the difference (subtract) between the accepted value and the experimental value, then divide by the accepted value. To determine if a value is precise find the average of your data, then subtract each measurement from it. This gives you a table of deviations. Then average the deviations.Gliffy tool used to find out the accuracy of algorithm.

Table 4.3 Accuracy of results

Dataset	**Size of Dataset**	**Decision Tree**	**Naive Bayes**
Laptops	Small(200 instances)	100%	92.857%
Comfort Sets	Medium(1500 instances)	99%	81.67%
Books	Large(4672 instances)	87.29%	63.24%

- **Time(in seconds):-**The number of (machine) instructions which a program executes during its running time is called its time complexity in computer science. This number depends primarily on the size of the program's input, that is approximately on the number of the strings to be sorted (and their length) and the algorithm used.Gliffy tool used to find out the time complexity of algorithm.

Table 4.4 Time Complexity

Dataset	**Size of Dataset**	**Time**	**Decision Tree**	**Naive Bayes**
Laptops	Small(200 instances)	To build mode To Test Model	0.41 sec 0.04 sec	0.68 sec 0.28 sec
Comfort Sets	Medium(1500 instances)	To build mode To Test Model	0.16 sec 0.14 sec	0.15 sec 0.31 sec
Books	Large(4672 instances)	To build mode To Test Model	0.32 sec 0.24 sec	0.45sec 0.28 sec

By above comparisons,it has been concluded that decision tree gives more fast and accurate result compare to naive bayes algorithm

V.APPLICATIONS

1.TRAVELLING

Travel recommender systems try to mimic the interactivity observed in traditional counselling sessions with travel agents when users search for advice on a possible holiday destination. From a technical viewpoint, they primarily use a content-based approach, in which the user expresses needs, benefits,and constraints using the offered language(attributes). The system then matches the user preferences with items in a catalog of destinations .

2. ENTERTAINMENT

Entertainment commendation system consider other things, such as, an entertainment item enjoying time since it could change over the course time.Personalized entertainment items recommendation is required to help millions of people narrow the universe of potential items to fit their unique taste.Finding a group of users based on the item he/she buys or provides feedback and then recommend popular items in the group.Recommending items to the producers such that they can entertain us more.

3. FOOD

All food recommender systems play a vital role in providing food items meeting preferences and adequate nutritional needs of users as well as persuading them to comply positive eating behaviors.Group recommendation functionalities are very useful in the food domain, especially when a group of users wants to have a dinner together at home or have a birthday party in a restaurant.

4.PROPERTY

Property recommender system can give recommendation to potential buyer after they declare their interest in one house on sale. These recommendation is made from a weight-set of some criteria, which summarize from questionnaire before.

V.SUMMARY

Recommendation systems help users discover items they might not have found by themselves and promote sales to potential customers, which provide an effective form of targeted marketing by creating a personalized shopping experience for each customer. Lots of companies have such kind of systems, especially for e-commerce companies like Amazon.com, an effective product recommendation system is very essential to their businesses. Inclusion of encryption technique to user’s password will provide security for our proposed system. They can be used to predict the rating for a product that a customer has never reviewed, based on the data of all other users and their ratings in the system. To examine and compare their effectiveness, we implement these three algorithms and test them on some existing datasets.

VI..REFERENCES

[1]Venkata Rajeev P & Amrita Vishwa Vidyapeetham,"Recommending products to customers using opinion mining Of online product Reviews & features",International Conference on Circuit,Power & Computing Technologies,2015

[2]Ming-Hsiung Ying,Yeh-Yen Hsu,"A Commodity Search System for Online Shopping Based on Ontology and Web Mining"IET The Institution of Engineering & Technologies,2014

[3]Miss Lovenika Kushwaha, Prof. Sunil Damodar Rathod,”New Opinion Mining Technique for Online Product Reviews and Features”,Multidisciplinary Journal of Research in Engineering and Technology, Volume 2, Issue 4, Pg.852-858

[4]Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7(1):76–80, 2003.

[5]http://jmcauley.ucsd.edu/data/amazon/links.html

[6]https://cseweb.ucsd.edu/~jmcauley//pdfs/sigir15.pdf

[7]Lars Backstrom and Jure Leskovec. Supervised random walks: Predicting and recommending links in social networks. Proceeding of WSDM 2011, pages 635–644, 2011.

EYE MOVEMENT AND VOICE BASED HUMAN COMPUTER INTERACTION

Sushil Singh, Bobby Bhatt, Shashank Chaubey, Raiyan Arab and Prof. Satishkumar Varma

Department of IT, PCE, New Panvel

Abstract*— Nowadays use of computer is the basic requirements needed by everyone for students to employees and children to elderly people, but it becomes difficult for a disabled person to use computers. This concept refers to the process of eye tracking and the movement of the eye and determining where the user is looking. The aim of this project is to present a technique used for eye tracking. Eye movements are both fast as an input device and are natural as a means of pointing when compared to other input devices. Users will typically look at the area of the screen where they wish to move to before they physically operate a mouse. However, a person's gaze is easily distracted by objects in the peripheral vision resulting in unwanted eye movements away from the object of interest. People don't necessarily think about the eye movements that they are making and so they are often performed subconsciously. But with practice, it is possible for a person to control their gaze such that it can be used as an effective computer input pointing device. The technique being used for face detection and eye detection is Viola Jones and for voice, Mel Frequency Correlation Coefficient (MFCC) technique is being used. The hardware by which we are going to present our concept is a Web Camera.

Keywords—Matlab, Application, Eye Movement, Voice Recognition, Face Detection, Disabled Person, Face Recognition, Retina, Eye Gaze, Mouse Control, HCI, Viola Jones, MFCC .

I. Introduction

Recently there has been a growing interest in developing natural interaction between human and computer. Several studies for human-computer interaction in universal computing are introduced. The vision-based interface technique extracts motion information without any high cost equipments from an input video image. Thus, vision-based approach is taken into account an effective technique to develop human computer interface systems. For vision-based human computer interaction, eye tracking is a hot issue. Eye tracking research is distinguished by the emergency of interactive applications. However, to develop a vision-based multimodal human computer interface system, an eye tracking and their recognition is done. Real time eye input has been used most frequently for disabled users, who can use only their eyes for input.

2. Literature Survey

2.1 Literature review

Measuring gaze point

Kristian Lukander describes a prototype system, which enables the measurement of gaze point on the screen surface of a hand-held mobile device, without constraining the user's natural movements. The method is software based, and integrates a commercial eye tracking device with a magnetic positional tracking device. The evaluation of the system shows that it is capable of producing valid data with adequate accuracy.**[1]**

Retina based Mouse Control

The retina tracking using mouse control tool is developed which works with a regular camera and can be used for several real time applications. By using your eyes instead of your mouse, you can select what you are looking at. This tool can be a big step for armless people using the computer.**[2]**

Eye gaze tracking

Yiu-ming Cheung has developed many potential attractive applications including human–computer interaction, virtual reality, and eye disease diagnosis. For example, it can help the disabled to control the computer effectively. In addition, it can support controlling the mouse pointer with one's eyes so that the user can speed up the selection of the focus point. Moreover, the integration of user's gaze and face information can improve the security of the existing access control systems.**[3]**

Detecting of eye gaze by Ba Linh Nguyen is used in a lot of human computer interaction applications. Most of them use intrusive techniques to estimate the gaze of a person. For example, user has to wear a headgear camera to fix the position of their eyes with the view of screen on the camera, or use an infrared light on camera to detect the eye. Here, it introduces a non intrusive approach which is very cheap solution to detect the eye gaze with a camera simple, user does not have to wear the headgear or using any expensive equipment.**[3]**

Eye movement based HCI

Using eye movement for controlling the computer, Ramsha Fatima [4] improves the experience of working with the computer as it is faster and gives the illusion that the computer is complying with the users' thought. It can be used either exclusively or in combination with other input technologies such as eye movement can be used along with a button so that it confirms the users' intentions for performing critical tasks and reduce the chances of error. It does not require any training and can thus be used by a layman. It can act as a boon for a person with motor disability as it does not require any motion but simple eye movements. It can give them a greater controlled over their surrounding and help them in interacting with the world.**[4]**

2.3 Literature Summary

Table 2.1 Summary of literature survey

SN	Paper	Remarks
1.	Kristian Lukander et al.2004[1]	A prototype system was defined and implemented for tracking the point of gaze on a mobile screen or user interface.
2.	Arslan Qamar Malik et al.2007[2]	In this paper, a working of the product has been described as to how it helps the special people share their knowledge with the world. Number of traditional techniques such as Head and Eye Movement Tracking Systems etc. exist for cursor control by making use of image processing in which light is the primary source.
3.	Ba Linh Nguyen et al.2009[3]	The problem of eye gaze tracking has been researched and developed for a long time. To track the eye gaze we have to deal with three principle problems: detecting the eye, tracking the eye and detecting the gaze of the eye on the screen where a user is looking at. In this paper we introduce the methods existed to solve these problems in the simple way and achieving high detection rate.
4.	Ramsha Fatima et al.2015[4]	In this paper we discussed various eye tracking techniques that can be used to find the line of gaze of the user. We then discussed some of the algorithms that can be used to implement these eye tracking techniques. The main aim of this paper is to propose new applications utilizing eye gaze that are suitable for standard user.

3.1 Overview

Controlling the computer mouse using the eyes movement requires a fast and effective algorithm, that's brought us to decrease the running time of the tool to the minimum by dividing the operation into few steps and using a tracking algorithm in order to avoid unnecessary calculations.

3.1.1 Existing System

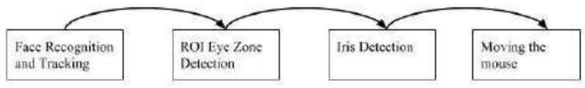


Fig. 3.1 Existing system architecture

1) Face Detection and Face Tracking

The first part is to detect the face zone and make a useful face tracking; we used the “Viola-Jones” algorithm in order to detect the face zone from the whole picture. The Viola Jones algorithm involves the sums of image pixels within rectangular areas along the image and also several iterations until the strong classifier is found where each iteration the distribution of weights of weak classifier is recalculate , therefore may take a long period to operate. Track the face location instead of applying the Viola-Jones algorithm for each frame, helps to avoid loss of time and unnecessary calculations. The “Viola-Jones” algorithm detects the initial face zone where, in the hue representation, the mass center is searched as shown in figure 3.1. The HSV (hue, saturation, and value) are common cylindrical-coordinate representations of points in an RGB color model, where the hue representation is the angle around the central vertical axis corresponds to each cylinder. We use tracking in order to find all the next face box locations. By finding the location the mass center of the face box, we can detect the new location of the face center. The face sizing may be changed at each frame; In order to adjust the face box size for the current frame we keep a stable white present at every face box.

2) ROI Eye Zone Detection

The second part receives a masked RGB frame in addition to the face box that was found in the previous step. From the gray scale version of the masked frame we pass to a logical picture, using an adjusting threshold. An appropriate threshold depends on many variables such as face orientations, position of the webcam, external illumination interference, brightness of the eyes zone etc. Therefore using an adjusting threshold is very important for the success of this tool [4]. The desirable threshold region should be defined once by the user at every significant change of the lightning. In order to decrease running time, the function receives the threshold of the previous frame and defines it as the initial new threshold, that’s decrease the times the function is calculating an appropriate threshold. In order to find eye zone we apply morphologic operations, first "opening" and after "closing" on the logical image. In order to find the edge of the eye zone, we find the gradient image which is show in Figure 3.1. An image gradient is a directional change in the intensity or color in an image. In our logical image it shows the existing edges in the image. The y coordinate of the left-up corner of the eye box is found by the summing all the pixels in every y line, and taking the y that have the maximum sum. We assumed y reach from the middle of the face box width to the upper border. We choose the eye zone's size and the x coordinate of the left-up corner to be proportional to the face box size and location. In order to have the iris into the eye zone the user has to keep the eyes approximately in a straight line.

3) Iris Detection

After the eye zone detected, we need to pick the two irises. To avoid lightning disturbance we use an adjusting threshold to create a binary version from the gray scale image and the threshold received from the last image in order to reduce the calculation. We used morphologic operations for reduce the noise disturbance and to create islands there the irises are related. At this point we used connected-component labeling and filter all the labels that are smaller than the two biggest connected component labels. Next we found the center of mass of each component and by a constant shifting we mark the center of iris zone in each eye.

4) Mouse Control

After detecting the iris of both eyes, the mouse move will be by the head shift or by the iris move when the head is steady. We detect a head shifting and changing the cursor respectively when both eyes are moving to the same direction. When the head is not moving we move the cursor by the iris looking. In each eye area, we find the labels after we move to a binary image using an adjusting threshold. We noticed that when the eye is looking for one size, the binary image have more white on the other side. We eliminate labels that closed to the floor of the image, which appear due to the bright skin area under the eye. For the noiseless image we subtract the number of white on the right half with the one on the left half. If in both eyes the derivative acting the same, we determine by the result the iris direction and the cursor shifts. In addition, left click will operate when the user close his eye for only 2 seconds. Right click will operate when the user close his eye for only 5 seconds. When the user closes his eyes, the binary images become pure black.

3.1.2 Proposed System

Image Processing

a. The web camera captures the face image and given as an input. The face image is then detected and verified by using the Viola Jones Face Detection Algorithm. Then after detecting the face, it is processed for matching it with the stored images in the system. The matching is done with the Correlation Algorithm, which is a process of extracting information from the images as shown in figure 3.2. The further process is carried out only the after face matches with stored images.

b. At an interval of five seconds Viola Jones algorithm is run and the position of retina is captured and as the retina is getting moved, the cursor movement is done.

c. Similarly, the blink is captured and if placed on any folder the folder gets opened or the specific task is performed.

Speech Recognition

d. While the Image Processing is being done, If I say “Change” or any other predefined word, then the system gets shifted from Image Processing to Speech Recognition.

e. Then within another five seconds if responded with the predefined command the comparison is done using MFCC algorithm and verified and if matched the given action is performed.

f. If not responded within 5 sec then again it returns to image processing.

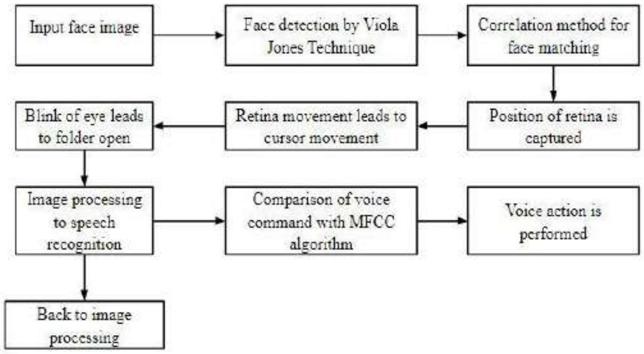

Input face image
Face detection by Viola Jones Technique
Correlation method for face matching
Position of retina is captured
Retina movement leads to cursor movement
Blink of eye leads to folder open
Image processing to speech recognition
Comparison of voice command with MFCC algorithm
Voice action is performed
Back to image processing

Fig. 3.2 Proposed system architecture

3.2 Implementation Details

The implementation detail is given in this section.

3.2.1 Viola Jones Algorithm

Viola-Jones technique is based on exploring the input image by means of sub window capable of detecting features. This window is scaled to detect faces of different sizes in the image. Viola Jones developed a scale invariant detector which runs through the image many times, each time with different size. Being scale invariant, the detector requires same number of calculations regardless of the size of the image.The system architecture of Viola Jones is based on a cascade of detectors. The first stages consist of simple detectors which eliminates only those windows which do not contain faces. In the following stages the complexity of detectors are increased to analysis the features in more detail. A face is detected only if it is observed through the entire cascade. These detectors are constructed from integral image and Haar like features shown in figure 3.3.

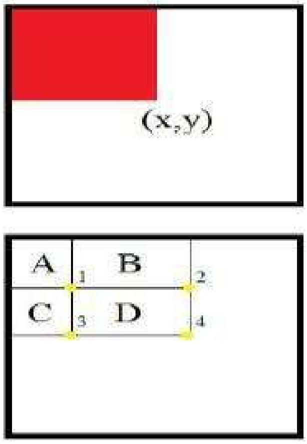


Figure 3.3. Viola Jones integral image construction.

The first step of this algorithm is to convert the input image into an integral image. This is done by making each pixel equal to the entire sum of all pixels above and to the left of the concerned pixel. By doing so, sum of all pixels inside any given rectangle can be calculated using only four values.

Sum of the rectangle, ABCD = D - (B + C) + A...(3.1)

The face detector in Viola Jones method analyzes a sub-window using features. These features consist of two or more rectangles. Each feature gives a single resultant value which is calculated by subtracting the sum of the white rectangle(s) from the sum of the black rectangle(s). Different types of features are shown below.

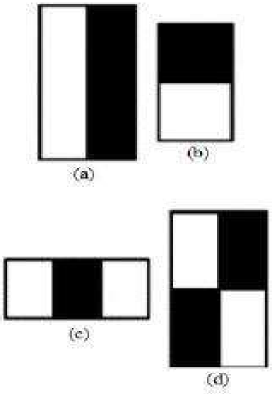


Figure 3.4. Viola Jones haar like features.

Viola and Jones used a simple classifier built from computationally efficient features using Ada Boost for feature selection. Ada Boost is a machine learning boosting algorithm that constructs a strong classifier through a weighted combination of weak classifiers. Mathematical description of weak classifier is,

Where x is a sub-window, f is the applied feature, p the polarity and θ is threshold that concludes whether x should be classified as a negative (non-face) or a positive (face).

Viola-Jones face detection algorithm scans the detector several times through the same image – each time with a new size. The detector detects the non face area in an image and discards that area which results in detection of face area. To discard non face area Viola Jones take advantage of cascading. When a sub window is applied to cascading stages, each stage concludes whether the sub window is a face object or not. Sub windows which contain some percentage of having faces are passed to next stage and those which are not faces are discarded. Final stage is considered to have a high percentage of face objects.

3.2.2 Correlation Algorithm

Correlation is a measure of the degree to which two variables agree, not necessary in actual value but in general behavior. The two variables are the corresponding pixel values in two images, template and source. Cross Correlation is used for template matching or pattern recognition. Template can be considered a sub-image from the reference image, and the image can be considered as a sensed image. The matching process moves the template image to all possible positions in a larger source image and computes a numerical index that indicates how well the template matches the image in that position. Match is done on a pixel-by-pixel basis as shown in figure 3.5.

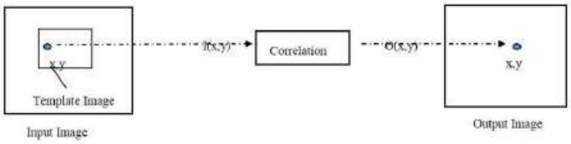


Figure 3.5 Method of matching the image using correlation.

3.2.3 Mel-Frequency Cepstral Coefficient (MFCC)

Mel Frequency Cepstral Coefficients algorithm is a technique which takes voice sample as inputs. After processing, it calculates coefficients unique to a particular sample. The simplicity of the procedure for implementation of MFCC makes it most preferred technique for voice recognition.

1) Voiced/Unvoiced Detection

Pre-processed signals are estimated for their energy and then weighted using the Dyadic Wavelet Transform (DTW) on each 256 samples/frame. The lowest energy level is at scale $\delta_1 = 2^1$ and the highest energy level is $\delta_5 = 2^5$.

Segments of sound signal with its largest energy level estimated at scale $\delta_1 = 2^1$ are therefore identified as unvoiced segments, otherwise found to be voiced segments. The following equation is the energy threshold defining as unvoiced segment;

$$uv = (n \mid \delta_i = 2^1); n= 1, \ldots., N \ (1)...(3.2)$$

At witch is the unvoiced segment of the n segment with energy at scale maximized.

2) Acoustic Feature Extraction

Only voiced segments of speech signal are processed for MFCC extraction. The procedure to determine MFCC is described as follows:

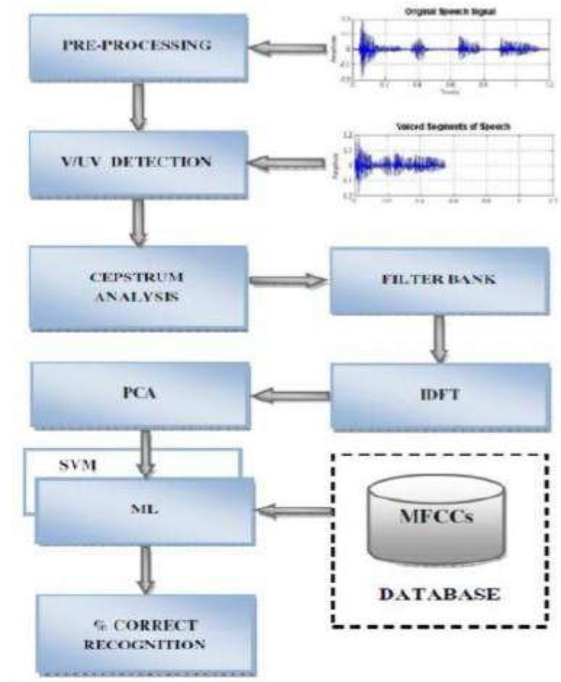


Figure 3.6. Work flow for MFCC based speech classification.[4]

3.3 Evaluation Parameters

No.	Function or Module	Estimated Cost
1	GUI	130
2	Database Operations	200
3	Back end	350
4	UI	170
	Total Cost	850

Table 3.1 LOC cost

F	Complexity Factors	Value Ratings out of 5
1	Are there distributed processing functions?	2
2	Is Performance Critical?	4
3	Is the code designed to be reusable?	3
4	Will the system run in heavily utilized OS?	3
5	Is the internal processing complex?	4
		Total= 16

Table 3.2 Complexity Factors

Acknowledgment

It is our privilege to express our sincerest regards to our project guide Prof. Satishkumar Varma for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of our project.

We deeply express our sincere thanks to our principal Dr. Sandeep M. Joshi and our Head of the Department Prof.Sharvari Govilkar for encouraging and allowing us to present the project on the topic "**Eye Movement and Voice Based Human Computer Interaction**" at our department premises for the partial fulfillments of the requirements leading to the award of B-tech degree.

We take this opportunity to thank all our lecturers who have directly or indirectly helped for steering us through as well as easy phases of the project in a result oriented manner with concern attention and support.

I would also thank all my colleagues and other people involved in this project for their constant help and support.

References

1. *Kristian Lukander,* "Mobile Usability - Measuring gaze point on hand-held devices", CHI '04 Extended Abstracts on Human Factors in Computing Systems,Pages 1550-1556, Austria, 2004.
2. *Arslan Qamar Malik and Jehanzeb Ahmed,* "Retina Based Mouse Control (RBMC)", World Academy of Science, Engineering and Technology, Volume 1 , Number 7, 2007.
3. *Ba Linh Nguyen*, "Eye Gaze Tracking", International Conference On Computing and Communications Technologies, pp.1-4, 2009.
4. *Yiu-ming Chang and Qinmu Peng*, "Eye Gaze Tracking With a Web Camera in a Desktop Environment", IEEE Transactions on Human-Machine Systems, Volume 45, Number 4, pp.419-430, Aug 2015.
5. *Ramsha Fatima, Atiya Usmani and Zainab Zaheer*, "Eye movement based human computer interaction", 23rd International Conference on Recent Advances in Information Technology (RAIT), pp. 489-494, 2016.

IOT Based Trustworthy Parking System

Ragil, Rajesh, Simran ,Veena ,Prof K.S. Charumati

Bachelor of Information Technology

Pillai College Of Engineering

Abstract: **With the ever increasing number of automobiles in metro cities, there is an acute shortage of parking space to fullfill the demand. Even for small durations paid parking facilities are charging high parking fees. Building and providing a digital parking system for the customers is focused by the Digital Parking Management System. The parking system will be for two wheelers as well as four wheelers. This paper demonstrates the implementation of the GPS based parking system using Google API for locating the nearest free parking spot. While registration, user has to provide RC book and Aadhar card details. For the authentication ,user has to undergo ethical question test through which it will be decided whether to register the user to the app or not. When the customer comes at the parking gate the OTP is checked which is send when the booking is done via android and slot is allotted to the user. By this the details of the vehicle are taken updated in the database. The system will then check if there is space available in the parking area and accordingly grant access to the customer. When the parking is full no vehicle will be allowed. The space available in the parking lot will be continuously updated in the system so that the entry of the vehicle can be controlled. The parking space for two wheelers and four wheelers will be different.**

Keywords-Car parking,Security questions,Trust,Verification.

I. Introduction

As the technical foundation of smart parking system, computing devices (e.g., smart phones, wireless sensors and personal laptops) turn progressively smaller, cheaper and more powerful.Finding a parking space is a common challenge faced by millions of citizens every day. A location-based application could help to this user with this problem as it would guide him depending on his current location. Global Positioning System (GPS) is a widely used technology for this purpose and it is constantly being improved. Let's imagine a driver who arrives to a shopping center looking for the place to park his car. Let's also imagine that the shopping center is on sale and therefore it is bursting with people. If the user needs to buy something quickly, something that he forgot the previous day when he did his weekly shopping, and he is also in a hurry because he just quit from his job for a few minutes, he would need extra help to find the best parking-position. The driver is not concerned with the shopping center entrances that are far away from his current location, rather he wants to choose one from several entrances near his current location and, if possible, closer to the requested shop. Personal costs will be reduced considerably using this technology.We have added recharge module therefore user has to register into the system and he will get message of balance on his mobile. It will be avoiding ticket-jamming problems for the ticket processing machines as well.Entry-point and exit-point will be handled in a fast manner without having to stop the cars so that traffic jam problem will be avoided during these processes.Vehicle owners will not have to make any payments at each Entry-point thus a faster traffic flow will be possible.Since there won't be any waiting during Entry-point and exit-points the pollution problem will be avoided.Automated parking system certainly reduces the total cost of parking system infrastructure without re-modifying the existed hardware.

One of the challenging problems for many vehicle owners in big cities is where to park their vehicles. If the parking slot is known in advance one can save precious time and fuel wastage. In our proposed system the user is informed about the parking slot availability at a particular parking location

Literature Survey:

Various methods are present for development of intelligent parking systems. We mainly focus on a management system that assists drivers to find parking spaces in a nearest parking community, and satisfies the needs of both parking providers and drivers. We now introduce several existing parking guidance approaches and show their limitations. We simulate the parking system performance under different parking management strategies

LiteratureReview:Vision based car parking system : A vision based car parking system was developed which uses two types of images (positive and negative) to detect free parking slot. In this method, the object classifier detects the required object within the input. Positive images contain the images of cars from various angles. Negative images do not contain any cars in them. The coordinates of parking lots specified are used as input to detect the presence of cars in the region. However, limitations may occur with this system with respect to the type of camera used. Also, the co-ordinate system used selects specific parking locations and thus camera has to be at a fixed location. Limited set of positive and negative images may put limitations on the system.

Number Plate Recognition technique : For developing autonomous car parking system uses image processing basis to process the number plates of the vehicles. In this system, the image of the license number plate of the vehicle is acquired. It is further segmented to obtain individual characters in the number plate. Ultrasonic sensors are used to detect free-parking slots. Then the images of number plate are taken and analyzed. Simultaneously, the current timing is noted so as to calculate the parking fees. The LCD displays 'FULL' sign to indicate that a parking slot is not available. However some limitations with the system include background color being compulsorily black and character color white. Also, analysis is limited to number plates with just one row.

Car Parking using Image Processing : In this system, a brown rounded image on the parking slot is captured using camera and processed to detect the free parking slot. The information about the currently available parking spaces is displayed on the 7-segment display. Initially, the image of parking slots with brown-rounded image is taken. The image is segmented to create binary images.

The noise is removed from this image and the object boundaries are identified. The image detection module determines which objects are round, by determining each object's area and perimeter. Accordingly, the free parking space is allocated.

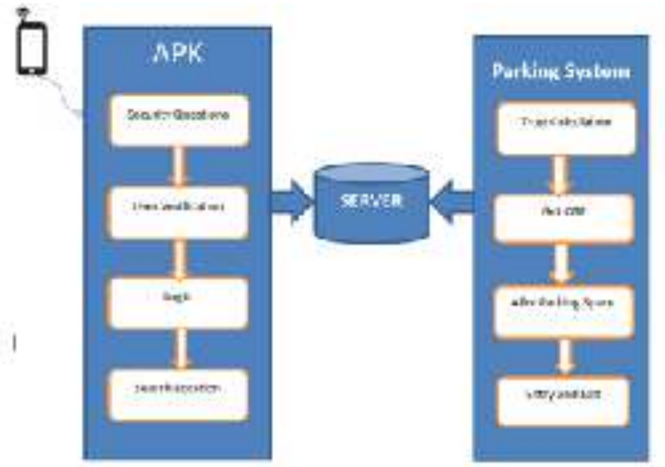
In this Parking Communities have been presented and provide trust management without any central authority. Vehicles create communities ,trusted groups helping their members to find parking in the community.Algorithms used are encryption and signature as well as mathematical trust model. There may be a DOS attack when too many vehicles come together to park at the same time. Here trust is important so whether the user should trust the information received about the free parking spots and ignore other free spots on the way.

Proposed System Architecture

Android

The user has to first get registered to the app through which the user can search various car parking areas nearby and can book the slot. For getting registered, user has to pass the ethical question set which will determine if the user should be registered for the application use. After getting registered the user can book the car parking slot and also search for the nearby locations.

The proposed system makes use of the concept of crowdsourcing to enable users to tag free or paid parking spots in their vicinity and other areas using GPS. It will enable the users to provide the starting and ending coordinates in the form of latitude and longitude of the area. This spot will be assigned a unique id and stored in the database along with its coordinates.



- It helps the visitors in finding out the availability of a parking slot, get the availability confirmed.
- It helps the parking owner to monitor the vacant slot availability
- The proposed plan saves the time of visitors in searching and booking a parking slot.
- The tedious job of parking owner to allocate the vacant slot in a methodical and organized manner is simplified as visitor himself chooses the suitable

parking place for his vehicle and the process is made more efficient

System Design

This system has been designed as a set of independent modules that communicate with each other through the use of Android

- M1 - Communications: This module keeps track of the state of the parking spaces.The users who wants to park his vehicle will register through android app and admin will verify this user at his location through the OTP.
- M2 - GPS: It will help find locations to park the vehicle through the areas where we want to park the vehicle
- M3 - GUI: It manages user registration and displays the slots for parking vehicles and tell the slots which are available and which are filled.
- M4 - Outside Parking Manager: This module controls the parking system in the particular location giving parking slots to the registered users
- M5 - Configuration: It manages the GPS configuration,Android and Desktop App Configuration.

The system relies heavily on a Mysql database that stores all the user data, request data, and data about the parking spots. The user table will store the details about each user. No information like age, sex, etc will be stored. Only the username will be stored to make sure that the parking spots being marked are marked only by humans. The request data table will store all the request from users of the system. This storing of requests will help us mine the data and find areas for which the requests are the most. We can then publish these results to show the areas where there is a need for parking infrastructure. The requests will be time-stamped so that we can create visualizations of request data with recognizable patterns at different times of the day, month, year etc. The parking spot data will be stored in a separate table and will store the latitude and longitude of the parking spots along with the username of the user who submitted that spot. Each parking spot will have a rating. The people who use the system will rate the parking spot based on their experience. As a result, the parking spot with the highest rating is the most genuine one. Similarly, the parking spots with a low rating can be inferred to be "risky". The users can also report bogus parking spots marked by users. The bogus parking spots will be eventually removed from the system. Every Android based smart-phone is equipped with a GPS chip. By using the GPS facility we can get the coordinates of the user. Using these coordinates we search for free parking spots nearest to these coordinates using Maps API.

System Requirements

All possible requirements of the system to be developed are captured in this phase. Requirements are set of functionalities and constraints that the end-user (who will be using the system) expects from the system. The requirements are gathered from the end-user by consultation, these requirements are analyzed for their validity and the possibility of incorporating the requirements in the system to be development is also studied. Finally, a Requirement Specification document is created which serves the purpose of guideline

Performance requirement:

Some Performance requirements identified is listed below:

- The database shall be able to accommodate a minimum of 10,000 records of customer
- The software shall support use of multiple users at a time for storage.
- There are no other specific performance requirements that will affect development.

Safety requirement:

The Application may get crashed at any certain time due to virus or operating system failure. Therefore, it is required to take the application backup.

Security requirement:

Application will allow users to access the system. There are TWO types of users namely Administrator and Customers. Security is based upon the individual user ID and Password. Some of the factors that are identified to protect the software from accidental or malicious access, use, modification, destruction, or disclosure are described below. Keep specific log or history data sets

- Check data integrity for critical variables
- Assign certain functions to different modules

- Restrict communications between some areas of the program
- Check data integrity for critical variables
- Communication needs to be restricted when the application is validating the user or license.

Hardware and Software Specifications

User Interfaces: The interface used in GUI must be easy to understand. This interface serves as a bridge between the user and the software. It also makes the user interaction with the system easy.

The user interface includes:

- Screen formats /Organizations: The introductory screen will be the first to be displayed which allows the user to log in using their id and password.
- Windows formats /Organizations: When the user chooses a particular topic then the information pertaining to that topic will be displayed in a new window, which will allow multiple windows to be available on the screen, and the user can switch between them.
- Data Format: The data entered by the user will be alphanumeric.
- End Message: When there are some exceptions, error messages will be displayed promptly by the user to re-enter the details when an event has taken place successfully.

Hardware interfaces: The system must basically support certain hardware and these must be an interface between them.

Communication interfaces:

This web application keeps track of every user dealing with the application. The communication is established with the help of a web application. With appropriate algorithm, Digital car parking system is introduced.

Evaluation metrics

The quality of a domain system can be evaluated by comparing recommendations to a test set of known user ratings. These systems are typical measured using precision and recall.

Precision: A measure of user ratings given by the community people, determines the relevant user retrieved out of all users retrieved. Precision (P). It is the proportion of recommended users those are actually good

Recall: A measure of completeness, determines the retrieved user out of all relevant user. It is the proportion of all good user recommended.

Summary

In various researches of Smart parking systems, different authors implemented numerous systems which have dynamic arrangement scheme for helping in different needs of drivers and service providers, which are based on real-time parking information however, as indicated in the tables of merits and demerits in this paper, more innovation is still needed to clear the gap as far as SPS is concerned.

References

[1] Hamada R.H. Al-Absi Patrick Sebastian Justin Dinesh Daniel Devaraj Yap Vooi Voon, “Vision-based automated parking system.”, 10th International Conference on Information Science, Signal Processing and their Applications, 2010.

[2] M.O. Reze M.F. Ismail A.A. Rokoni M.A.R. Sarkar, “Smart parking system with image processing facility”, I.J. Intelligent Systems and Applications, 2012.

[3] R. Yusnita Fariza Norbaya Norazwinawati Bashruddin, “Intelligent parking space detection system based on image processing”, International Journal of Innovation, Management and Technology, 2012.

[4] Julian Timpner & Lars Wolf "Trustworthy Parking Communities: Helping Your Neighbour To Find A Space " IEEE Transactions On Dependable and Secure Computing,Vol-13, 2016

[5] M. Fengsheng Yang, Android Application Development Revelation, China Machine Press, 2010.

[6] Shoup, D., “Cruising for parking”. Transport Policy, 2006

[7] J. Sherly, D. Somasundareswari “Internet of Things Based Smart Transportation Systems” International Research Journal of Engineering and Technology (IRJET) Volume: 02 Issue: 07 Oct-2015

[8] J. Dongjiu Geng, Yue Suo, Yu Chen, Jun Wen, Yongqing Lu, Remote Access and Control System Based on Android Mobil Phone”, Journal of Computer Applications, 2011.

[9] Pallavi Mane , Radha Deoghare , Samiksha Nagmote , Shubhangi Musle , Shraddha Sarwade “Android based Smart Parking System” International Journal of Innovative Research in Computer and Communication Engineering ,Vol. 3, Issue 5, May 2015

[10] J. Wolff, T. Heuer, H. Gao, M. Weinmann, S. Voit and U. Hartmann, "Parking monitor system based on magnetic field sensors," IEEE Conf. Intelligent Transportation Systems, 2006.

[11] C.Laugier and F.Thierry, "Sensor-based control architecture for a car-like vehicle", International Conference on Intelligent Robots and Systems, 1998.

[12] Hongwei Wang and Wenbo He, “Reservation-based SPS” The first international workshop on cyber-physical networking systems, Dept .Computer, Electrical Eng, University of Nebraska-Lincoln, NE, USA, 978-1-4244-9920-5/11. IEEE, 2011.

Enhancement of Security in Mobile Banking Applications Using Two factor Authentication

Sahil Sudhakaran[1], Ruchita Phalak[1], Rishabh Gupta[1], Shivani Raja[1] and Dr. Madhumita Chatterjee[2]

1 students, 2 Faculty, PCE, Department of Information Technology

***Abstract*—**

Mobile banking applications are increasingly being used for performing basic banking operations like transfer of money or viewing account details. These applications store important, personal and financial information of the user. While using these applications the major concern of any user is that 'is this secure'.If such confidential details are compromised the user might have to face heavy financial loss.

In this project, we are developing mobile banking application which includes two layers of security. First level includes OTP encryption and the second layer is biometrics. The basic idea behind using two factor authentication is to provide additional layer of security which makes it difficult for the attacker to gain access to user's personal information. The proposed system will provide an increased security in mobile banking by making use of combination of two factor authenticationwith SIM/IMEI verification.

Key words: Insecure network, Mobile Security, Mobile Banking, Two level security,

I. Introduction

1.1 Mobile Security

The confidential data like personal, financial, business etc. are now stored on smart phones. It is made possible due to features provided by banking application. Thus, mobile device security has become increasingly important in mobile computing.

The attacker targets these smart phones and exploits its inherent weaknesses. These attacks can come from the communication mode such as SMS (Short Message service), MMS (Multimedia Messaging Service), WIFI, etc. Security measures are deployed at different layers or levels as a counter measure to such attacks. The implementation of security measures must be observed at every level from designing to development of the operating system, software layers, and downloadable apps. The basic purpose is to securely transfer the information to the actual user.

1.2 Types of Security Threats

In mobile technology the three major areas where malicious attacks are possible are- the network the device and the data centre. The threats in each area are:

The Device: Sensitive data storage, no encryption/weak encryption, dynamic runtime injections, unintended permissions.

The network: insecure Wi-Fi, packet sniffing, session hacking, man-in-the-middle.

The data centre: weak input validation, Brute force attack, SQL injection.

1.3Two Factor Authentication:

In Two-factor authentication (2FA), the user needs to pass two layers of security before he can claim his identity. Once the system validates the user's identity, he is granted access to the system.

Mechanisms can be such as,

Knowledge (something they and only they know),

Possession (something they and only they have),

and Inherence (something they and only they are).

Example, Withdrawal of money from ATM. The user must know ATM pin-code (Something user knows) and must have bank card (Something user posses). Only the combination of the above two can lead to successful transaction.

Mobile-phone two-step authentication is more secure than single-factor password protection but suffers some security concerns. Phones can be cloned and apps can run on several phones; cell-phone maintenance personnel can read SMS texts. Not least, cell phones can be compromised in general, meaning the phone is no longer something you and only you have.

To avoid such issues, 2FA was introduced. In order to authenticate themselves user can use their personal access codes and one time valid dynamic password (OTP). The pass code is forwarded to their mobile device by SMS or push notification. In all three cases, the advantage of using a mobile phone is that there is no need for an additional dedicated token, as users tend to carry their mobile devices around at all times.

II. Existing System

In the existing approach, mobile includes security features based only on identity-based access control techniques such as username and password. These features are less secure for financial data. As data is transferred via an insecure network, identity based technique can be easily hacked by the attackers. Due to this, most of the users don’t prefer to use mobile devices for financial transaction. Another set-back of the current system is that, the user’s financial data can be easily accessed from any other mobile device other than user’s device. Thus by knowing the user’s username and password, attacker can easily gain access to the account from another device.

The first step towards secure mobile banking operation starts by creating a secure session between the mobile device and the bank server. The secure session is created using TLS handshake Protocol. The TLS encrypts the data that is exchanged between the client and server which helps in protection against the intruders. After this we can start with authentication phase. In the first level of authentication the client is authenticated using a username and password. Then in the second level of authentication, the user’s mobile device is authenticated through IMEI and SIM serial number. The IMEI and the SIM serial number are field in the background without user interference. All these four credentials, i.e. username, password, IMEI and SIM serial number are combined into array and sent to the server side. The parameters in this array are then checked one by one by the server. If all the parameters are correct then the user is granted access to his/her financial account, otherwise he/she will be asked to enter the username and password again

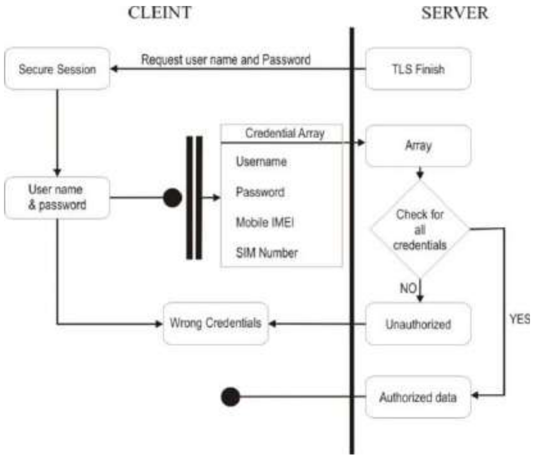


Fig. 1

III. Proposed System

The application is built in android studio. The app provides the basic banking processes such as transfer of money, view account balance etc. Our primary focus is the security and not the application.

Considering the current scenario, we have retained the data transmission protocol to be TLS(operates on TCP/IP). The user credentials will be obtained such as user id, password, IMEI SIM number. It will be validated at the server end. Along with this, we are incorporating Two Factor Authentication Technique. One is OTP encryption and the other one is Biometric Fingerprint Scanning. We will be creating different layers of security.

3.1 User Authentication Block:

Firstly, the user is authenticated using USER ID and PASSWORD. When the application is started by the user, he is asked to enter his credentials. User ID and Password is sent to the server database. Database is searched for a match of the same user id and the password. If the match is found the user has to authenticate the device which is the second module.

3.2 Device Authentication Module:

We have used the concept of IMEI/SIM number. This step basically, informs the system that the device being used to login or register is blacklisted or not. If the device is blacklisted, the user is denied access. The application fetches the IMEI number and the SIM serial number when the user registers for the first time. This fetched information is stored in the backend and whenever the user logs in, every time the IMEI/SIM number is verified, only then the user is granted access. Thus, allowing user to access the bank account only from the registered device.

3.3 Scanner Detection :

Our primary target is smart phones with scanner. But in case we encounter a device without a scanner, we need to have an alternative method. This process is just to identify whether the device has a scanner or not. In devices with inbuilt finger print scanner, biometrics will be used for authentication before the transaction. For devices without scanner we have included a set of security questions to authenticate the user.

3.4 Two Factor Authentication Module:

This module is only applicable if any transaction of money is involved. Basic activities like viewing account or transaction history can be done after user and device authentication. This is to reduce the load on the server.

Here we have provided two layers of security.
OTP Encryption
Biometric Finger Print Scanning

So**, OTP** is basically, one time password, valid for a very short period of time. But still it has some threats like the device may be used by some other unintended user or session hijacking,then the OTP can be misused. Thus, to cover these disadvantages, we are implementing OTP encryption.
We have used AES algorithm with 128 bit key size. Since AES is a symmetric key algorithm, key is generated at the server end and the user end (i.e in the app). The keyis generated by a combination of IMEInumber, SIM numberand the registration time. These three parameters are put together into a variable and then a substring of 32 digit in length is generated by shuffling which serves as the key. Each time a different key is generated due to shuffling. This will make difficult for the attacker to generate his own key to decrypt the OTP message. The OTP will get encrypted at server, then transferred to the user. The app will decrypt it and automatically fetch the OTP number.

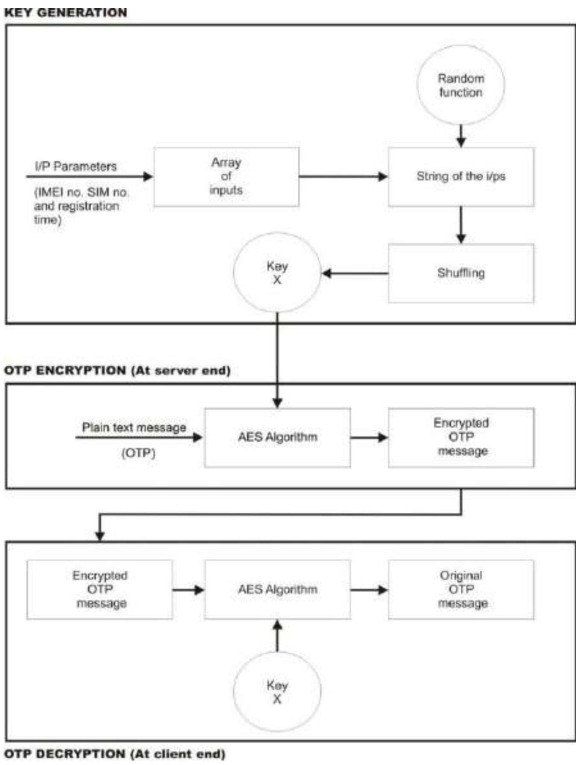


Fig. 2 Key generation block diagram

Next, factor is **Biometrics**. We have implemented finger print scanning as a biometric security system for our application. For the simplicity of creating this project, we will be using built-in function of fingerprint scanning. The already registered fingerprints in the mobile device will be used. The app does not provide separate registration for fingerprint. This can be implemented in the future.

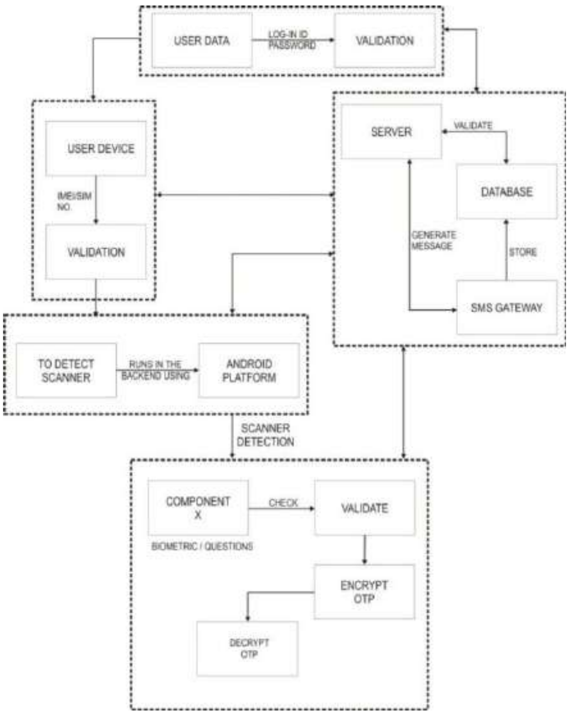


Fig.3 System Architecture

I. Results

The basic purpose of using the OTP is to check whether the user is using his own device or not. The Normal OTP sent over a channel can be easily intercepted by Man in the Middle Attack. If the attacker can get access to OTP, he can perform every function of an authorized user. Thus by using the encrypted OTP, even if attacker is able to get access to the OTP, he will not be able to use it. The attacker needs to generate the key to decrypt the message, before he can manipulate the message.

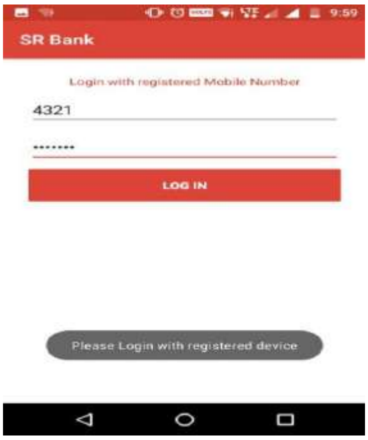
The idea about improvising on the existing security measures, we were able to overcome many of the disadvantages or shortcomings, from our approach. We are supporting this statement in the form of test cases given below.

4.1 Test case Scenario

Scenario 1: The user is trying to login

Case 1: The user enters the correct User-id and Password using the registered mobile device i.e. the SIM and IMEI number is registered, this will allow the user to login successfully.

Case 2: IF user enters any of the credentials wrong or uses an unregistered device, access will be denied.



Please Login with registered device

Login from unregistered device

Scenario 2: The user wants to transfer the money

Case 1: If the user enters the correct holder name and the account no. and wants to transfers the amount 500 to a registered user i.e. the name is in the beneficiary list, then he can successfully transfer the amount.

Case 2: If the user wants to transfer amount but he enters the wrong user holder name or the account no. Or if tries to transfer amount more than 25000 (Maximum limit), then he cannot transfer the money successfully.

Scenario 3: The user loses his mobile to an attacker

The authorized users A and B are registered users.

Case 1: The user C (attacker) can try to login using an authorized device.
If the attacker changes the SIM then the attacker would not be able to login as the device is authenticated using SIM+IMEI, thus the stored SIM+IMEI values would not be matched.

Case 2: The user C (attacker) can try to login using an authorized device.
If the attacker forges the login-id and password of an authorized user then he would be able to login as the SIM+IMEI are matched with the login credentials.
But at the later stages, the attacker would not be able to continue as an authorized user.

Scenario 4: The user has changed his device

Case 1: The user will not be able to login using his own credentials as the IMEI number will get changed and normal users avoid changing IMEI number as it involves high risks. The user can re-register by setting new login-id and password.

Case 2: If the user wants to login using the same credentials then the user needs to contact the respective bank as only the bank can unregister the user and allow him to re-register using the old login-id and password.

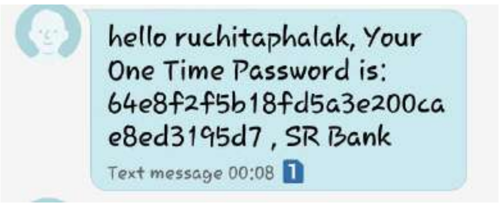
Scenario 5: The attacker tries to forge the IMEI number to a registered IMEI number

Consider that an attacker knows the login-id and password and attacker tries to forge (Change) his own device IMEI number to the registered IMEI number.

Case 1: This attempt will block the attacker from using the application because the device is registered using IMEI + SIM number. Even if IMEI number is changed, the application will not recognize the device as IMEI and SIM will not match.

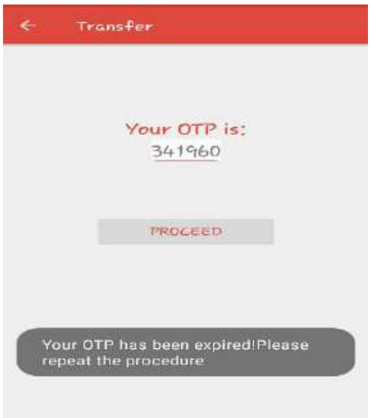
Scenario 6: In case of session hijacking

Case 1: In case the attacker imposes session hijacking attack while a user has already logged in, he will fail. Since the OTP generated is being encrypted and it cannot be easily decrypted and its validity expires within 2 minutes.



hello ruchitaphalak, Your One Time Password is: 64e8f2f5b18fd5a3e200cae8ed3195d7 , SR Bank

OTP in SMS inbox



Your OTP is:
341960

Your OTP has been expired!Please repeat the procedure

OTP has expired

II. Conclusion

Thus, by analysing the results of our project, we can conclude that, Two factor authentication improves the security of the application. Having multilayered security architecture enhances the security, makes it more reliable and gives better data protection. Our implementation also retained the processing capability of the device. No additional third party apps or devices are needed.

There still remains scope of further improvement. As we said our focus is on scanner based devices, we can think of creating new security measures for the scanner-less devices.

We can use mechanisms such as email verification, we can set limit on transaction amount. If someone tries to carry out transaction beyond that, the user will be notified by an SMS or an Email.

In this project, we have used built in finger print scanner. It can be improved by actually implementing image processing concepts. Further we can use other forms of biometrics i.e voice recognition, face detection etc.

III. References

[1]Enhancement of Security in Mobile Banking Applicationsby Mr. Mayur Waghmare1 Ms. PriyaGolekar Mr. AkshayHatwar Ms. RushaliParimal Ms. Akanksha Hiware,Department of Computer Science & Engineering RTMN University, India.1st January, 2017

[2]Issues and Security Measures of Mobile Banking Apps by Sameer Hayikader , Farah NurafiqahHanisbinti Abd Hadi, Jamaludin Ibrahim. 1st January, 2016

[3]OTP based two factor authentication using mobile phone by Mohamed HamdyEldefrawy, Khaled Alghathbar, Muhammad Khan. 12th July, 2011

[4]Security issues in biometric authentication by QinghanXio.

[5]A Proposal to Improve Security of Mobile Banking Applications
by M. Elkhodr, S. Shahrestani, K. Kourouche.
11th January, 2013